Missingness Adapted Group Informed Clustered (MAGIC)-LASSO: A novel paradigm for prediction in data with widespread non-random missingness.

Amanda E. Gentry*[1], Robert M. Kirkpatrick[1], Roseann E. Peterson[1], Bradley T. Webb[1,2]

1. Virginia Institute for Psychiatric and Behavioral Genetics, Department of Psychiatry, Virginia Commonwealth University, Richmond, VA, USA.
2. GenOmics, Bioinformatics, and Translational Research Center, Biostatistics and Epidemiology Division, RTI International, Research Triangle Park, NC, USA.

Corresponding Author
Amanda Elswick Gentry, Ph.D.
Post-doctoral Fellow
Virginia Commonwealth University
Virginia Institute for Psychiatric and Behavioral Genetics, Department of Psychiatry
Amanda.gentry@vcuhealth.org
804.855.4703
ORCID: 0000-0002-6425-9340

Robert M. Kirkpatrick, PhD
Assistant Professor
Virginia Commonwealth University
Virginia Institute for Psychiatric and Behavioral Genetics, Department of Psychiatry
Robert.Kirkpatrick@vcuhealth.org

Roseann E. Peterson, Ph.D.
Assistant Professor
Virginia Commonwealth University
Virginia Institute for Psychiatric and Behavioral Genetics, Department of Psychiatry
Peterson.Roseann@gmail.com
ORCID: 0000-0001-6402-849X

Bradley Todd Webb, Ph.D.
Omics Research Scientist, Principal Investigator
RTI International
GenOmics, Bioinformatics, and Translational Research Center
Biostatistics and Epidemiology Division
bwebb@rti.org
ORCID: 0000-0002-0576-5366

Abstract:
The availability of large-scale biobanks linking rich phenotypes and biological measures are a powerful opportunity for scientific discovery. However, real-world collections frequently have extensive non-random missing data. Machine learning methods are able to predict missing data but performance is significantly impaired by block-wise missingness inherent to many biobanks. To address this, we developed Missingness Adapted Group-wise Informed Clustered LASSO (MAGIC-LASSO) which performs hierarchical clustering of variables based on missingness followed by sequential Group LASSO within clusters. Variables are pre-filtered for missingness and balance between training and target sets with final models built using stepwise inclusion of features ranked by completeness. This research has been conducted using the UK Biobank (n>500k) to predict unmeasured Alcohol Use Disorders Identification Test (AUDIT.) The phenotypic correlation between measured and predicted total score was 0.67 while genetic correlations between independent subjects was >0.86, demonstrating the method has significant accuracy and utility.

Introduction:

Biobanks are large-scale high-dimensional collections of biomedical information offering significant opportunities for scientific discovery, with many collections containing thousands of data points on tens of thousands of individuals. Many biobanks collect biospecimens and perform genome-wide assessments of genetic variation and increasingly other omic measures such as gene expression, epigenetic modifications, and proteomics which allow comprehensive agnostic investigations of the relationships between complex human diseases and traits with genetic and environmental influences. These powerful resources are increasingly accessible to the larger scientific community facilitating novel investigations and discovery. The breadth of phenotypes in biobanks represents an opportunity for machine learning (ML) approaches to further discover unexpected relationships complementing directed *a priori* hypothesis testing. However, the scale of biobanks also presents challenges including significant missing data, much of which is non-random.

There is a growing list of available biobanks for scientific discovery including the UK Biobank[1] (UKB) which has enrolled over half a million UK residents, all of whom provided biological samples for genotyping. Volunteers in the UKB also provided access to their electronic health records, hospitalization records, biological samples, and answers to survey questions regarding diet, lifestyle habits, and mental health; phenotypic measures available to link with genetic measures total in the thousands. In the US, the National Institutes of Health is funding the All of Us[2] biobank effort, which has enrolled nearly 25% of its goal of one million participants who will provide biological samples, genotypic data, electronic health records, and answers to several series of survey questions. Similarly, BioBank Japan[3] has sampled over 200,000 participants with one of 47 common diseases and collected genetic information along with health records and other phenotypic information. Many additional biobanks are currently available to researchers and construction of new biobanks continues, motivated in part by the necessity of collecting large sample sizes to study the genetics of complex traits.

Structural characteristics in biobanks present challenges for data analysis. Many biobanks do not administer every test or survey to each participant, as budget considerations, for example, often dictate how many participants receive more costly testing, such as imaging. In order to mitigate dropout and participant fatigue, a subset of questionnaires may be sent to each participant; requests for participation in a particular survey may have been sent to a portion of subjects and only a subset of those were returned. Similarly, subsets of subjects may be chosen to participate in additional surveys according to previous responses, where the decision logic for these selections may not be clear or available to researchers. These practices, while pragmatic for cost and volunteer retention, may result in widespread, block-wise missingness across the full biobank, in which large subsets of the full sample have completely missing values for a portion of question categories. This missingness is non-random across the full set of available measures in the biobank such that no subset of subjects with complete information exists in the sample.

Missingness patterns can severely limit the power for epidemiological and genetic analyses of any single trait. Traditionally, data missingness can be addressed through imputation where a missing-at-random structure can be reasonably assumed. Some commonly used approaches include *k* nearest neighbors[4] or Multivariate Imputation by Chained Equations[5] (MICE). These methods borrow information across the available data to infer missing points, but because biobank missingness is generally pervasive across all phenotypes and often decidedly non-random, these traditional imputation methods are not appropriate for filling in the missing values. Where imputation is inappropriate, row-wise deletion is sometimes employed to drop

subjects who have missing observations. However, given the block-wise nature of biobank missingness, this sort of deletion can render the dataset many orders of magnitude smaller. Given the phenotypic depth of the biobank, there is an opportunity to apply ML methods to leverage the existing data to predict missing values. Utilization of ML, or "data mining," as it is often called, has continued to rise across many applications including genetics. For example, the PsychENCODE project[6] employed deep learning techniques to predict functional ramifications in the brain of genome-wide association study (GWAS) hits associated with psychiatric disorders. Advances in technology and cloud resources continue to ease the computational burden of applying ML methods to high-dimensional genomic data with many of the machine learning methods themselves based in statistical techniques long established theoretically and proven empirically[7].

A subset of ML approaches have been adapted to account for some level of predictor missingness and applied to missing phenotype imputation. MI-LASSO[8], for example, integrates Multiple Imputation (MI) of missing predictors with the Least Absolute Shrinkage and Selection Operator (LASSO) for a hybrid approach applicable where missingness may be assumed to be random. PhenIMP[9] and extensions[10] use related phenotypes to impute a difficult to collect phenotype in order to boost power. While PhenIMP can impute using only summary information from other phenotypes, it relies on distributional assumptions which make the approach impractical where many phenotypes are categorical and do not conform to such assumptions. Similarly, the PHENIX[11] method was designed to impute missing phenotypes in a Bayesian framework in the presence of other informative data but also requires distributional assumptions and does not drop non-informative input measures, thereby prohibiting variable selection. Other approaches developed by Yuan et al.[12] and expanded upon by Xiang et al.[13] specifically addresses block-wise missingness structures with a focus on imputing entire blocks of missing data, specifically where neuroimaging data is present. While innovative and effective for applications involving a small number of well-defined blocks of data, this method is not applicable to the structure of large-scale data wherein the blockwise missingness patterns are highly inconsistent across subjects and the number of blocks is large.

Given the variety of available ML approaches and characteristics of biobanks, there is significant need for an ML solution for imputing missing phenotypes which collectively (1) is capable of including categorical and/or non-normally distributed predictors, (2) produces interpretable models, (3) incorporates penalization or variable selection such that it could be generalizable, and most importantly, (4) is applicable and robust in the presence of non-random, blockwise missingness. While many traditional ML methods could satisfy the first three interests, most are intolerant to missingness in the predictors, precluding out-of-the-box application of available methods.

As a proof of principle, we selected the UKB to serve as an example application of our proposed ML method. The data freeze (UKB Application 30782, approval date Sep 3, 2018, using data baskets created Sep 28, 2019 and May 20, 2019) contained 9,613 phenotypes on 502,536 subjects. We chose the Alcohol Use Disorders Identification Test (AUDIT) survey from UKB as our target outcome which was directly measured in 157,162 (31.2%) participants. Here, we describe a novel ML approach and demonstrate its usefulness in leveraging thousands of measured phenotypes in order to predict an unknown, unmeasured phenotype and show how this predicted outcome boosts power for downstream analyses including GWAS and cross-trait genetic correlation studies.

Results:

MAGIC-LASSO:

We developed an adaptation of the Group Least Absolute Shrinkage and Selection Operator (Group-LASSO) machine learning method for penalized regression to address the shortcomings of existing, software-implemented ML methods for predicting phenotypes in the presence of non-random, blockwise missingness named the Missingness Adapted Group Informed Clustered (MAGIC)-LASSO.

LASSO background:

As a member of the family of penalized regression ML techniques, the LASSO[14] is well established and popular. Often presented in the context of the elastic net[15] formulation, the general formula for the LASSO may be found under the linear regression paradigm by estimating the values which minimize:

$$\frac{1}{2N}\sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2 + \lambda\left[\frac{(1-\alpha)\|\beta\|_2^2}{2} + \alpha\|\beta\|_1\right],$$

where there are $i = 1, \dots, N$ total participants, y_i represents outcome observation i and x_i^T is a vector of predictors. The λ is a tuning parameter that adjusts the amount of shrinkage (penalization) applied to the model and α mixes the L_1 and L_2 penalties, where:

$$L_1\ penalty, \|\beta\|_1 = \sum_{i=1}^{N}|x_i|, \text{ and}$$
$$L_2\ penalty, \|\beta\|_2^2 = \sum_{i=1}^{N}|x_i|^2,$$

such that when $\alpha = 1$, the formula becomes the LASSO and when $\alpha = 0$, the formula becomes ridge regression[16]; mixing the two parameters using $0 < \alpha < 1$ results in the classic elastic net design. The nature of the ridge penalty prevents any coefficient estimate to shrink to exactly zero, making it more useful for addressing multicollinearity than dimension reduction, while the LASSO encourages sparsity and the coefficients of some covariates are allowed to shrink to exactly zero, making it a useful tool for variable selection. The mixing parameter is generally chosen by the user and set for the duration of the experiment. Where the overarching goal is to identify a parsimonious set of covariates from a large pool which may accurately approximate an outcome, setting $\alpha = 1$ is generally appropriate because the LASSO is well suited to achieve this. The tuning parameter (λ) is best chosen through cross-validation, a method by which a single observation is held out during the fitting process and the resulting model used to predict the outcome of the left-out observation. Repeating this process for every observation in the dataset allows for the calculation of a mean error rate of prediction. The lambda associated with the model with the smallest mean error rate of prediction is generally the best model[7].

Group LASSO background:

In traditional LASSO, categorical covariates can be included by coding them according to a numerical scale. This is not ideal, as it assumes equal spacing between and inherent ordering within categories[17]. A solution to this is to dummy-code the k-level categorical measures by augmenting the representation of a single variable into $k - 1$ binary variables. However, traditional LASSO treats each of these as individual measures which may result in shrinkage of some but not all categories within a single covariate, rendering interpretation of selected variables difficult. The Group LASSO[18] was developed to encourage sparsity at the factor level, such that all categories of a given variable are included or excluded from the model as a set. The Group LASSO is found as the solution to:

$$\frac{1}{2}\left\|Y - \sum_{j=1}^{J} X_i\beta_j\right\|^2 + \lambda\sum_{j=1}^{J}\left\|\beta_j\right\|_{K_j},$$

$$\text{Where } \left\|\beta_j\right\|_{K_j} = \left(\beta_j^T K_j \beta_j\right)^{1/2}.$$

In the simplest form, K_j can be an identity matrix, but it may take a variety of forms. To fit a model within this framework, each categorical variable must be expanded into binary dummy-coded columns. Our MAGIC-LASSO approach utilizes the conventional Group-LASSO fitting algorithm, but applies it in an innovative, iterative manner in order to overcome the challenges of the blockwise missingness design.

MAGIC-LASSO Overview:

In brief, MAGIC-LASSO procedure involves (1) Characterizing missingness, (2) filtering variables for general missingness and for balance across training and target sets, (3) variable clustering based on missingness, (4) iterative Group-LASSO and variable selection within clusters, and (5) cross cluster model building with variables prioritized by informativeness. **Figure 1** describes the flow logic of the MAGIC-LASSO.

Characterizing missingness and general filtering: This first step is to create a subset of variables suitable for downstream investigation. This includes removing potential predictor variables that are (a) excessively sparse (>80% missingness), (b) categorical with excess, sparse levels such as ICD codes in a collapsed matrix format, (c) unstructured where the format is inappropriate for modeling, such as free text, date values, or array variables, or (d) invariant. After initial filtering, we identify and remove variables for which missingness patterns were highly skewed between prediction and training sets for the outcome of interest. Due to blockwise missingness, there may be variables which pass the first filtering step but are not informative in the target dataset. In other words, where data completeness is highly correlated with the variable of interest. This is not to be confused with correlation among the phenotypic measures themselves, which is generally not of concern since the LASSO procedure more capable of handling many measures with varying degrees of collinearity than traditional linear regression[7].

General background on ML training and test sets:

This filtering step relies on the identification of a so-called measured set, also referred to here as a training set, the subset of subjects with the primary outcome measured, and an unmeasured set, or a prediction set, the remaining subjects for whom the outcome of interest was unmeasured and for whom the ML procedure will predict the missing phenotype. **Figure 2** illustrates an example of how a dataset may be subdivided into these measured and unmeasured sets.

Balancing:

When training an ML model to predict unmeasured variables, the learning occurs on the subset of data for which complete observations are available, i.e., the measured, or training set and is then implemented in the unmeasured, or the prediction set. The algorithm learns how to predict unobserved data by modeling patterns that exist in observed data. Where certain variables are largely measured in conjunction with the primary outcome of interest in the training set but are largely unmeasured in the prediction set, an ML algorithm which relies on these measures for prediction will perform poorly, since the inputs will be largely missing.

For a given experiment, partition the total number of observations into those in the measured and unmeasured sets $N_{measured} + N_{unmeasured} = N_{total}$ and for each additional phenotype k, quantify $n_{k,measured}$ and $n_{k,unmeasured}$ the number of observations present in $N_{measured}$ and $N_{unmeasured}$, respectively. Then calculate a filtering parameter:

$$\tau_k = \frac{n_{k,measured}/N_{measured}}{n_{k,unmeasured}/N_{unmeasured}},$$

where τ_k represents the ratio of the proportion of observations present in the measured set to the proportion of observations present in the unmeasured set, for phenotype k. Plotting is helpful to determine empirically a useful cutoff for $\tau = t$.

Clustering:

The Group-LASSO, like many ML procedures, cannot accommodate missing data and relies on row-wise deletion of observations where one or more variables are missing. One strategy to mitigate reduction in sample size from requiring complete information across all covariates is to segregate the variables into blocks according to patterns of missingness and apply the ML procedure within that subset of measures. In the MAGIC clustering step, variables are grouped to minimize missingness while maximizing sample sizes in order to optimize downstream within-cluster prediction performance. First, pairwise observation counts for every pair of phenotypic variables are calculated. Using this pairwise count matrix, we calculated the Euclidean distance of these measures to feed into an average-linkage agglomerative hierarchical clustering procedure to discover the inherent groupings of variables based on missingness. This clustering procedure begins with each variable in its own cluster and proceeds by combining two clusters for each step until all variables reside in a single cluster. The clustered tree may be cut at some point to obtain the clustering assignments. Exact height for cutting is determined empirically by examining mean observation count per variable in the cluster, number of variables in the cluster, and the number of complete cases for that subset of measures.

Iterative Group-LASSO:

The cut tree provides groups of variables within which the complete data observation count is maximized. Limiting each cluster to only the complete data therefore, the Group-LASSO is applied to each cluster individually. Each model utilizes k-fold cross-validation to choose the penalty term with minimal prediction error and variables in each model were retained if they achieved non-zero effect estimates, where k is chosen to be small and approaching n, with consideration of computation resources. After applying the Group-Lasso to each cluster, variables retained by each model are aggregated across clusters, as illustrated in **Figure 3**. The set of aggregated, retained variables is then re-clustered using the same hierarchical clustering procedure and the Group-Lasso applied to each cluster. With each successive iteration of the clustering and Group-Lasso application, the phenotype space shrinks as the measures most predictive of the outcome are retained across iterations and the less-informative measures are dropped. **Figure 1** illustrates the flow of the algorithm, which continues until moderate parsimony is achieved.

Cross cluster model building:

Once the iterative procedure is halted, with p remaining phenotypes, the Group-Lasso is fit up to $(p-1)$ times using a stepwise procedure which orders the phenotypes according missingness. Beginning with the phenotype with least missingness and adding an additional phenotype each round, the Group-Lasso model is fit to the complete data on those phenotypes and k-fold cross-validation is used to determine the phenotypic correlation between the observed and predicted outcome. With each step, an additional phenotype is added and the Group-Lasso fit, and the procedure continues until every set has been fit, or there are no longer any complete cases in the successive set. The set of phenotypes producing the most predictive model is chosen as the final model.

Although each iteration of the Group-Lasso application is fit using cross-validation, it is optimal to further utilize a hold-out test set during the construction of the final model in order to rigorously assess performance. The proportion of the data assigned to the hold-out test set

depends on the size of the data set itself, although a hold-out set containing 10-30% of the data is typical[7]. Phenotypic predictive performance is assessed by plotting observed versus predicted observations in the test set and reporting the correlation between the observed and predicted sets.

Application to real world biobank data:

As a proof of principle, we applied MAGIC-LASSO to predict AUDIT in the UKB. The AUDIT is a ten-item, self-administered screening instrument for alcohol problems containing three questions surveying consumption and seven items surveying problems related to alcohol which comprise the AUDIT-C and -P subscales[19,20]. The AUDIT survey was part of the mental health battery of questionnaires and was returned by 157,348 UKB participants. Median total AUDIT score was 4, median AUDIT-C was also 4 , and median AUDIT-P was 0.

To construct a set of variables to be used in the predictive algorithm, we filtered the full set of 9,603 available phenotypes (not including the AUDIT measures) to remove measures (a) with fewer than 100,000 observations, (b) which were ICD codes, (c) which were unstructured, (d) which were invariant, or (5) which were repeated and measured at later longitudinal timepoints, such that only baseline measures were retained. **Figure 4** shows the sample sizes remaining after each filtering step. After these filtering steps, 631 curated, so-called top-level (i.e., baseline) variables remained. Further filtering for balance between the measured and unmeasured sets removed 277 more variables for which missingness patterns were highly skewed between prediction and training sets. Measures with a ratio of missingness in the predicted versus the training set of $t \leq 0.7$ were filtered out, leaving 354 variables.

Clustering the post filtered phenotype set resulted in an initial 12 clusters (**Table 1**). One cluster of 5 phenotypes was dropped because there were no complete cases in the cluster. Using 5-fold cross-validation, the first application of the Group-Lasso resulted in an aggregated total of 99, 106, and 123 phenotypes were retained across all clusters for the AUDIT-Total, AUDIT-C, and AUDIT-P, respectively. In the second iteration, phenotypes were grouped in 5, 6, and 4 clusters for AUDIT-Total, AUDIT-C, and AUDIT-P respectively and applying the Group-Lasso to each cluster resulted in an aggregate of 65, 80, and 54 phenotypes retained across the clusters for AUDIT-Total, AUDIT-C, and AUDIT-P, respectively.

Having reduced the phenotypic space by nearly a quarter for each score, the iterative Group-Lasso process halted. We then ordered the phenotypes in each set according to missingness and applied the Group-Lasso procedure to the set of phenotypes constructed in a forward stepwise manner, beginning with the phenotype with least missingness. **Table 2** shows the number of subjects with complete data, with the addition of each phenotype, including the breakdown of complete cases in the measured and unmeasured sets, as well as the phenotypic correlation from a predictive model constructed using each successive set of phenotypes. The phenotypic correlations and the ratio of proportion of complete cases from the measured and unmeasured sets are shown for each outcome in **Figure 5.** The stepwise procedure showed final models with 30, 18, and 20 input variables resulted in the best prediction for Total, Consumption, and Problems respectively. The final models resulted in 27, 18, and 20 non-zero coefficient estimates and test-set phenotypic correlations of 0.64, 0.71, and 0.48 for Total, Consumption, and Problems respectively.

Using the full measured sets for which both observed and predicted AUDIT scores were available, the phenotypic correlations were 0.65, 0.70, and 0.46 for Total, Consumption, and Problems, respectively. **Figure 6** shows the density curves of observed and predicted (both measured and unmeasured) for all three scores. Density curves of the prediction residuals are shown in **Figure 7**.

One significant advantage to evaluating ML methods including MAGIC-LASSO in biobanks such as UKB is the availability of genetic information on all subjects and methods to estimate SNP-based heritabilities (h^2) and genetic correlations (r_g). To explore the accuracy of the predicted phenotypes and evaluate their utility in downstream genetic studies, we estimated h^2 in each set and r_g between (a) observed and predicted in the subjects with measured AUDIT and (b) predicted AUDIT in subjects with and without direct measurement. We note that the last sets are completely independent with no information being shared in the model building step except for missingness balance.

Heritability: Within-subject GCTA based h^2 for AUDIT-T in men and women showed similar estimates between measured and predicted (range 0.089 – 0.139) (**Table 3)** and was similar to LDSC based estimates (range 0.047 – 0.087) derived from GWAS summary statistics. Of note, the estimated heritabilities of the predicted score are close to those of the observed score in both the measured and unmeasured sets. Using LDSC, we estimated heritabilities for men and women combined across the three outcome sets, see **Table 4 and Figure 8**. The LDSC estimates are slightly lower than the GCTA estimates, as expected. Heritability in observed and predicted AUDIT-Total and AUDIT-C are similar, while the point estimate for observed AUDIT-P is lower than that in predicted AUDIT-P.

Genetic Correlations: Using GCTA and only subjects with measured AUDIT, the r_g between the observed and predicted AUDIT-T was 0.863 (se 0.040) in men and 0.884 (se 0.032) in women. The LDSC-estimated r_g (**Table 5 and Figure 9)** between observed and predicted AUDIT in the measured set provides an indicator of prediction performance, with r_g between these sets of 0.919 (AUDIT-Total), 0.858 (AUDIT-C), and 0.792 (AUDIT-P.)

Discussion:

The goal of this methodological work was to develop an ML procedure which could predict missing phenotypes (1) accurately, (2) in an interpretable manner, and (3) in a generalizable framework, (4) using existing software, and (5) for application in biobank-scale datasets with block-wise missing data structures. Our novel MAGIC-LASSO approach achieves these goals, as demonstrated through the prediction of the AUDIT measures in the UK Biobank study. The consistently high phenotypic and genetic correlations across the observed and predicted sets indicates that the ML procedure is capable of predicting the missing phenotype with high accuracy and in a manner which faithfully reflects the underlying genetic contribution to the phenotype. It is further noteworthy to mention that in ML practice, predictive performance in the full set often overestimates the real-life potential of the algorithm to predict missing values. However, our predictive performance in the full set was nearly identical to that in the hold-out test set in the UKB AUDIT application, with a difference in phenotypic correlation of no more than 0.02 between the full and hold-out sets in all three AUDIT measures.

Prediction was less accurate in the AUDIT-P outcome as compared to AUDIT-Total and AUDIT-C. This demonstrates two considerations, first, that the distribution of the outcome can affect its prediction. Where observations are highly skewed and less evenly distributed across the potential range, prediction is rendered more difficult. Second, prediction performance varies based on the phenotype and the dataset, as observed with the AUDIT measures. The available phenotypes in UKB, in aggregate, lend better information to the prediction of AUDIT-Total and AUDIT-C than of AUDIT-P, although expansion of the phenotypes entering the MAGIC-LASSO model may improve the prediction of AUDIT-P. Furthermore, **Figure 5** demonstrates the differing architecture of the predicted scores in Total and Consumption versus Problems,

where the first few variables comprise the bulk of the prediction for AUDIT-T and AUDIT-C, while the prediction of AUDIT-P is composed of more variables of small effect.

Strengths of the MAGIC-LASSO include, first, it can be applied using existing packages in the R software environment. Second, the prediction process is straightforward and transparent. The MAGIC-LASSO is built on the foundation of the Group-Lasso, a statistically rigorous framework with well-established properties which allow the user access to the regression structure of the prediction. Third, it is applicable to large biobank-scale environments where missing-at-random structures cannot be assumed. The application of the MAGIC-LASSO for phenotype imputation can confer great power gains for genetic analyses, as demonstrated using AUDIT in UKB. AUDIT and genotypic data were directly measured in 117,559 European ancestry individuals in the UKB sample. Predicting AUDIT in the unmeasured subjects added 242,421 independent samples for downstream GWAS, representing a 56% increase in effective sample size. Finally, the MAGIC-LASSO is a flexible framework allowing for straightforward adaptations for application to datasets of various structures and outcomes of different characteristics.

Limitations of the MAGIC-LASSO framework include the limitations of the Group-Lasso procedure to account for interaction effects of covariates. The current demonstrated implementation of the approach is also limited to the linear regression framework. Penalized non-linear regression algorithms which can account for grouped covariates exist for the logistic regression framework[21] but not for the ordinal or polytomous outcome scenario, rendering application of the MAGIC-LASSO to item-level AUDIT responses, for example, not possible using existing software.

Despite these limitations, the method demonstrated strong predictive performance in the real data UKB application and represents an innovative contribution to the field of biomedical research in biobanks. The method is accessible through open-source software and transparent in nature, allowing the user to assess performance and understand the full regression procedure constructing the predicted outcomes. The MAGIC-LASSO is an additional tool now available to researchers to further harness the discovery potential inherent in large data collections and maximize the return on the financial and altruistic participant time and effort contributions invested in the assembly and management of biobank resources.

Methods:

Data management and application of the MAGIC-LASSO was conducted in in R[22] (v3.5.2) using packages *Matrix*[23] (v1.2.17), *fastDummies*[24] (v1.5.0), and *grpreg*[25] (v3.2.1) Clustering was conducted using *hclust* UPGMA method in base R and the Group-Lasso was fit using the *cv.grpreg* function in the *grpreg* package.

GWAS: To assess how well the predicted AUDIT outcome captures the underlying genetic factors influencing AUDIT, we calculate the heritability of observed and predicted AUDIT as well as the genetic correlations (r_g) between the observed and predicted outcomes. To this end, we conducted GWAS (bgenie[26] version 1.3) of the AUDIT-Total, AUDIT-C, and AUDIT-P scores in the measured and the combined measured plus unmeasured sets. We utilized the Neale lab GWAS filtering criteria[27] for MAF < 0.5%, INFO < 0.8, and HWE p-value < 10^(-6), adjusting for covariate effects of age, sex, and the first 20 PCs. The independent European subjects sample size for the GWAS was 359,980 with 117,559 and 242,421 subjects in the AUDIT measured and unmeasured sets, respectively.

Heritability and Genetic Correlation: GCTA[28] (version 1.93.2) was used to calculate heritabilities and the r_g between observed and predicted AUDIT, but only within the set of

participants on whom AUDIT was directly measured since the GCTA only allows r_g to be calculated across the same set of observations. We also utilized LDSC[29,30] (version 1.0.1) to estimate heritabilities and r_g between observed and predicted scores, both within the measured set and between the measured and unmeasured sets. Using LDSC, r_g can be estimated in independent samples by leveraging a reference set of genetic correlations (linkage disequilibrium) and GWAS test statistics.

Data Availability:

The UKB data utilized in this research is available to, "bona fide researchers for health-related research in the public interest[31]," through an application process accessible through the UKB website, https://www.ukbiobank.ac.uk/.

References:

1. 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 Med.* **12**, e1001779 (2015).
2. The All of Us Research Program Investigators. The "All of Us" Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).
3. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).
4. Troyanskaya, O. *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520–525 (2001).
5. van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software, Articles* **45**, 1–67 (2011).
6. Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, eaat8464 (2018).
7. Hastie, T., Tibshirani, R. & Friedman, J. *The elements of statistical learning*. (Springer, 2009).
8. Chen, Q. & Wang, S. Variable selection for multiply-imputed data with application to dioxin exposure study. *Stat. Med.* **32**, 3646–3659 (2013).
9. Hormozdiari, F. *et al.* Imputing Phenotypes for Genome-wide Association Studies. *Am. J. Hum. Genet.* **99**, 89–103 (07/2016).
10. Chen, Y., Peloso, G. M. & Dupuis, J. Evaluation of a phenotype imputation approach using GAW20 simulated data. *BMC Proc.* **12**, 56 (9/2018).

11. Dahl, A. *et al.* A multiple-phenotype imputation method for genetic studies. *Nat. Genet.* **48**, 466–472 (4/2016).

12. Yuan, L., Wang, Y., Thompson, P. M., Narayan, V. A. & Ye, J. Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. *Neuroimage* **61**, 622–632 (07/2012).

13. Xiang, S. *et al.* Bi-level multi-source learning for heterogeneous block-wise missing data. *Neuroimage* **102**, 192–206 (11/2014).

14. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* **58**, 267–288 (1996).

15. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.* **67**, 301–320 (04/2005).

16. Hoerl, A. E. & Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12**, 55–67 (02/1970).

17. Kutner, M. *Applied linear statistical models Michael H. Kutner ... [et al.]*. (McGraw-Hill Irwin, 2005).

18. Yuan, M. & Lin, Y. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Series B Stat. Methodol.* **68**, 49–67 (02/2006).

19. Higgins-Biddle, T. F., Saunders, J. C., Monteiro, J. B. & Babor, M. G. *AUDIT: The alcohol use disorders identification test: guidelines for use in primary care*. (Geneva World Health Organization, 2001).

20. Saunders, J. B., Aasland, O. G., Babor, T. F., De La Fuente, J. R. & Grant, M. Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO Collaborative Project on

Early Detection of Persons with Harmful Alcohol Consumption-II. *Addiction* **88**, 791–804 (06/1993).

21. Meier, L., Van De Geer, S. & Bühlmann, P. The group lasso for logistic regression. *J. R. Stat. Soc. Series B Stat. Methodol.* **70**, 53–71 (2008).
22. R Core Team. R: A Language and Environment for Statistical Computing. (2020).
23. Bates, D. & Maechler, M. Matrix: Sparse and Dense Matrix Classes and Methods. (2019).
24. Kaplan, J. fastDummies: Fast Creation of Dummy (Binary) Columns and Rows from Categorical Variables. (2020).
25. Breheny, P. & Huang, J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Stat. Comput.* **25**, 173–187 (3/2015).
26. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (10/2018).
27. Neale Lab GitHub. *GitHub* https://github.com/Nealelab/UK_Biobank_GWAS.
28. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
29. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
30. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
31. UK Biobank Website. https://www.ukbiobank.ac.uk/.

Acknowledgements:
This research has been conducted using the UK Biobank Resource application number 30782. AE Gentry, RM Kirkpatrick, RE Peterson, and BT Webb are supported by NIAAA P50AA022537. AE Gentry is supported by NIMH T32MH020030. RE Peterson is supported by NIMH K01MH113848 and The Brain & Behavior Research Foundation NARSAD grant 28632 P&S Fund.

Author Contributions: All authors contributed to the concept and design of the study. R.E.P. and B.T.W. contributed to the data acquisition. All authors contributed to the analyses, interpretation of the data and the results, as well as to substantive revisions of the manuscript.

Competing Interests Statement: The authors have no competing interests to disclose.

Table and Figure Legends:
Table 1: Clustering details for rounds 1 (a) and 2 (b). For each iterative round, number of phenotypes, complete cases per cluster, and number of phenotype measures retained after applying the Group-Lasso to each cluster are shown.
Table 2: Stepwise Group-Lasso application results for AUDIT-Total (a), AUDIT-Consumption (b), and AUDIT-Problems (c). For each step, the number of complete cases (subjects with complete data) in and the proportion of observations in the measured and unmeasured sets, as well as the phenotypic correlation for the prediction model built on that set, are shown. The highlighted line indicates the best performing model, with 30 phenotypes in AUDIT-Total, 18 in AUDIT-C, and 20 in AUDIT-P.
Table 3: GCTA estimated heritabilities (h^2) and genetic correlations (r_g) for AUDIT-Total. r_g calculated as the genetic correlation between the observed and predicted scores within the measured set only.
Table 4: LDSC GCTA estimated heritabilities (h^2) for observed and predicted AUDIT.
Table 5: LDSC estimated genetic correlations (r_g) between the observed and unobserved AUDIT in the measured and unmeasured sets.
Figure 1: Flow logic of the MAGIC-LASSO procedure. The MAGIC-LASSO procedure begins with filtering, followed by clustering, then iterative Group-LASSO application until parsimony is achieved.
Figure 2: Conceptualization of how a dataset may be subdivided into a measured and unmeasured set. Where N represents the full sample size, $N_{Unmeasured}$ and $N_{Measured}$ represent the subsets of subjects on whom the outcome of interest is either missing or measured, respectively. Then the amount of overlap in observation may be quantified for each of p additional variables.
Figure 3: Illustration of clustering and Group-LASSO procedures within the MAGIC-LASSO. The hierarchical clustering steps divides the phenotypic space into non-overlapping subsets based on missingness. The Group-LASSO is applied to each cluster and variables retained by the algorithm in each set are aggregated.
Figure 4: Flowchart of filtering in the UKB application set. The number of variables remaining after each filtering metric is applied is shown for the UKB data application example.
Figure 5: Stepwise model results. Phenotypic correlations and ratio of the proportion of total measured to unmeasured observations in the model set for (a) AUDIT-Total, (b) AUDIT-Consumption, and (c) AUDIT-Problems. Final selected model highlighted in blue.

Figure 6: Density curves of the observed and predicted scores. Outcomes in the observed and predicted in the measured and unmeasured sets plotted for (a) AUDIT-Total, (b) AUDIT-C, and (c) AUDIT-C.
Figure 7: Residual densities for AUDIT prediction. Density curves with means noted showing the distribution of the prediction residuals for (a) AUDIT-Total, (b) AUDIT-Consumption, and (c) AUDIT-Problems.
Figure 8: LDSC estimated heritabilities. SNP-based heritability estimates for the observed (green) and predicted in the measured (purple) and unmeasured (orange) sets for the AUDIT outcomes.
Figure 9: LDSC estimated genetic correlations. Genetic correlation estimated between the observed data and predicted scores in the measured sets (green,) the observed data and the predicted scores in the unmeasured sets (orange,) and the predicted scores in the measured and unmeasured sets (purple.)

Table 1: Clustering details for rounds 1 (a) and 2 (b). For each iterative round, number of phenotypes, complete cases per cluster, and number of phenotype measures retained after applying the Group-Lasso to each cluster are shown.

(a) Round 1:

Cluster	Number of Phenotypes	Number of Complete Cases	Phenotypes retained in AUDIT T	Phenotypes retained in AUDIT C	Phenotypes retained in AUDIT P
1	250	80839	28	24	52
2	12	19292	9	11	12
3	7	44941	7	7	7
4	6	16436	1	5	0
5	14	24888	9	8	8
6	2	27852	2	2	2
7	5	67766	5	5	5
8	5	0	-	-	-
9	11	225	5	6	5
10	31	45638	22	28	21
11	2	31086	2	1	2
12	9	120601	9	9	9

(b) Round 2:

AUDIT-T			
Cluster	Number of Phenotypes	Number of Complete Cases	Phenotypes Retained
1	46	16363	21
2	17	4031	13
3	7	12680	7
4	5	36004	5
5	24	27737	19
AUDIT-C			
Cluster	Number of Phenotypes	Number of Complete Cases	Phenotypes Retained
1	33	85210	17
2	19	5169	18
3	12	6312	8
4	7	12680	7
5	6	6608	5
6	29	33174	25
AUDIT-P			
Cluster	Number of Phenotypes	Number of Complete Cases	Phenotypes Retained
1	69	16922	14
2	19	8244	10
3	7	12680	7
4	28	6612	23

Table 2: Stepwise Group-Lasso application results for AUDIT-Total (a), AUDIT-Consumption (b), and AUDIT-Problems (c). For each step, the number of complete cases (subjects with complete data) in and the proportion of observations in the measured and unmeasured sets, as well as the phenotypic correlation for the prediction model built on that set, are shown. The highlighted line indicates the best performing model, with 30 phenotypes in AUDIT-Total, 18 in AUDIT-C, and 20 in AUDIT-P.

(a) AUDIT-Total

Number of Phenotypes	Number of Complete Cases	Number of Complete Cases (measured)	Number of Complete Cases (unmeasured)	Proportion of Measured Observations	Proportion of Unmeasured Observations	Ratio of Measured to Unmeasured	Phenotypic Correlation
1	502536	157162	345374	1	1	1	NA
2	502536	157162	345374	1	1	1	0.2684
3	501645	157089	344556	0.9995	0.9976	0.9981	0.3419
4	501639	157088	344551	0.9995	0.9976	0.9981	0.6044
5	501632	157088	344544	0.9995	0.9976	0.9981	0.6058
6	501515	157067	344448	0.9994	0.9973	0.9979	0.611
7	499725	156837	342888	0.9979	0.9928	0.9949	0.616
8	497760	156548	341212	0.9961	0.9879	0.9918	0.6081
9	489704	154505	335199	0.9831	0.9705	0.9872	0.6089
10	489654	154491	335163	0.983	0.9704	0.9872	0.6128
11	472443	149925	322518	0.954	0.9338	0.9789	0.613
12	453342	144398	308944	0.9188	0.8945	0.9736	0.6174
13	453339	144397	308942	0.9188	0.8945	0.9736	0.6187
14	433878	138101	295777	0.8787	0.8564	0.9746	0.6189
15	433378	137942	295436	0.8777	0.8554	0.9746	0.6197
16	432764	137767	294997	0.8766	0.8541	0.9744	0.6281
17	399391	131141	268250	0.8344	0.7767	0.9308	0.623
18	364427	125509	238918	0.7986	0.6918	0.8662	0.6035
19	331556	113971	217585	0.7252	0.63	0.8687	0.628
20	271230	95947	175283	0.6105	0.5075	0.8313	0.6155
21	215930	79510	136420	0.5059	0.395	0.7808	0.5544
22	163047	62004	101043	0.3945	0.2926	0.7416	0.5469
23	149450	56980	92470	0.3626	0.2677	0.7385	0.4953
24	149450	56980	92470	0.3626	0.2677	0.7385	0.5284
25	149450	56980	92470	0.3626	0.2677	0.7385	0.6179
26	149450	56980	92470	0.3626	0.2677	0.7385	0.6376
27	149450	56980	92470	0.3626	0.2677	0.7385	0.6376
28	91677	33330	58347	0.2121	0.1689	0.7966	0.6265
29	55671	19386	36285	0.1234	0.1051	0.8517	0.6366
30	26307	9862	16445	0.0628	0.0476	0.7588	0.639
31	26307	9862	16445	0.0628	0.0476	0.7588	0.637
32	26307	9862	16445	0.0628	0.0476	0.7588	0.637
33	26307	9862	16445	0.0628	0.0476	0.7588	0.637
34	12636	4904	7732	0.0312	0.0224	0.7175	0.6241
35	5844	2062	3782	0.0131	0.011	0.8346	0.6271
36	5844	2062	3782	0.0131	0.011	0.8346	0.6271
37	5844	2062	3782	0.0131	0.011	0.8346	0.6271
38	5844	2062	3782	0.0131	0.011	0.8346	0.6271
39	5844	2062	3782	0.0131	0.011	0.8346	0.6271
40	2612	852	1760	0.0054	0.0051	0.94	0.5728
41	0	0	0	0	0	NA	NA

(b) AUDIT-Consumption

Number of Phenotypes	Number of Complete Cases	Number of Complete Cases (measured)	Number of Complete Cases (unmeasured)	Proportion of Measured Observations	Proportion of Unmeasured Observations	Ratio of Measured to Unmeasured	Phenotypic Correlation
1	502536	157162	345374	1	1	1	NA
2	502536	157162	345374	1	1	1	0.2884
3	501645	157089	344556	0.9995	0.9976	0.9981	0.3471
4	501639	157088	344551	0.9995	0.9976	0.9981	0.6881
5	501522	157067	344455	0.9994	0.9973	0.9979	0.687
6	499529	156778	342751	0.9976	0.9924	0.9948	0.6881
7	496519	156224	340295	0.994	0.9853	0.9912	0.6922
8	488461	154161	334300	0.9809	0.9679	0.9868	0.6898
9	488411	154147	334264	0.9808	0.9678	0.9868	0.6901
10	475775	150955	324820	0.9605	0.9405	0.9792	0.6927
11	475775	150955	324820	0.9605	0.9405	0.9792	0.6935
12	461337	146357	314980	0.9312	0.912	0.9793	0.6896
13	461334	146356	314978	0.9312	0.912	0.9793	0.6932
14	442202	140174	302028	0.8919	0.8745	0.9805	0.6921
15	441698	140017	301681	0.8909	0.8735	0.9804	0.6979
16	441068	139835	301233	0.8898	0.8722	0.9803	0.6991
17	401599	127090	274509	0.8087	0.7948	0.9829	0.7018
18	332697	107054	225643	0.6812	0.6533	0.9591	0.7053
19	267747	89546	178201	0.5698	0.516	0.9056	0.6976
20	229101	77294	151807	0.4918	0.4395	0.8937	0.6938
21	181950	64355	117595	0.4095	0.3405	0.8315	0.6074
22	147982	53641	94341	0.3413	0.2732	0.8003	0.6106
23	109901	38096	71805	0.2424	0.2079	0.8577	0.5998
24	84491	28471	56020	0.1812	0.1622	0.8954	0.6134
25	67376	22807	44569	0.1451	0.129	0.8892	0.6026
26	50148	17732	32416	0.1128	0.0939	0.8319	0.6023
27	50148	17732	32416	0.1128	0.0939	0.8319	0.6021
28	46032	16357	29675	0.1041	0.0859	0.8256	0.5732
29	46032	16357	29675	0.1041	0.0859	0.8256	0.5975
30	46032	16357	29675	0.1041	0.0859	0.8256	0.6477
31	46032	16357	29675	0.1041	0.0859	0.8256	0.6609
32	46032	16357	29675	0.1041	0.0859	0.8256	0.6613
33	30054	8731	21323	0.0556	0.0617	1.1113	0.6677
34	25683	7353	18330	0.0468	0.0531	1.1344	0.6628
35	14602	4503	10099	0.0287	0.0292	1.0206	0.6663
36	9876	2979	6897	0.019	0.02	1.0535	0.6643
37	7380	2215	5165	0.0141	0.015	1.0611	0.6626
38	7380	2215	5165	0.0141	0.015	1.0611	0.6632
39	7380	2215	5165	0.0141	0.015	1.0611	0.663
40	7380	2215	5165	0.0141	0.015	1.0611	0.663
41	3311	1066	2245	0.0068	0.0065	0.9583	0.5482
42	1586	574	1012	0.0037	0.0029	0.8023	0.5793
43	760	256	504	0.0016	0.0015	0.8959	0.6511
44	0	0	0	0	0	NA	NA

(c) AUDIT-Problems

Number of Phenotypes	Number of Complete Cases	Number of Complete Cases (measured)	Number of Complete Cases (unmeasured)	Proportion of Measured Observations	Proportion of Unmeasured Observations	Ratio of Measured to Unmeasured	Phenotypic Correlation
1	502536	157162	345374	1	1	1	NA
2	501645	157089	344556	0.9995	0.9976	0.9981	0.2076
3	501639	157088	344551	0.9995	0.9976	0.9981	0.3199
4	501629	157088	344541	0.9995	0.9976	0.9981	0.3301
5	501512	157067	344445	0.9994	0.9973	0.9979	0.3386
6	499724	156837	342887	0.9979	0.9928	0.9949	0.3548
7	497759	156548	341211	0.9961	0.9879	0.9918	0.3438
8	489703	154505	335198	0.9831	0.9705	0.9872	0.3494
9	467267	148214	319053	0.9431	0.9238	0.9796	0.359
10	446935	141689	305246	0.9015	0.8838	0.9803	0.3706
11	446291	141504	304787	0.9004	0.8825	0.9801	0.3655
12	444751	141046	303705	0.8975	0.8794	0.9798	0.3718
13	357287	117944	239343	0.7505	0.693	0.9234	0.3797
14	280039	97110	182929	0.6179	0.5297	0.8572	0.3866
15	252356	88187	164169	0.5611	0.4753	0.8471	0.4084
16	252356	88187	164169	0.5611	0.4753	0.8471	0.4376
17	252356	88187	164169	0.5611	0.4753	0.8471	0.4569
18	252356	88187	164169	0.5611	0.4753	0.8471	0.4674
19	252356	88187	164169	0.5611	0.4753	0.8471	0.4685
20	147063	48890	98173	0.3111	0.2843	0.9138	0.4785
21	71542	26002	45540	0.1654	0.1319	0.797	0.4623
22	71542	26002	45540	0.1654	0.1319	0.797	0.4623
23	71542	26002	45540	0.1654	0.1319	0.797	0.4623
24	71542	26002	45540	0.1654	0.1319	0.797	0.4623
25	39805	13484	26321	0.0858	0.0762	0.8883	0.4663
26	39805	13484	26321	0.0858	0.0762	0.8883	0.4663
27	39805	13484	26321	0.0858	0.0762	0.8883	0.4663
28	39805	13484	26321	0.0858	0.0762	0.8883	0.4663
29	39805	13484	26321	0.0858	0.0762	0.8883	0.4663
30	17562	5429	12133	0.0345	0.0351	1.017	0.3982
31	0	0	0	0	0	NA	NA

Table 3: GCTA estimated heritabilities (h^2) and genetic correlations (r_g) for AUDIT-Total. r_g calculated as the genetic correlation between the observed and predicted scores within the measured set only.

Set	Outcome	Sex	h^2	SE (h^2)	N	r_g	SE (r_g)
Measured	Observed	Male	0.139	0.0108	50,912	0.863	0.040
Measured	Predicted	Male	0.091	0.0105	50,912		
Measured	Observed	Female	0.109	0.0087	64,768	0.884	0.032
Measured	Predicted	Female	0.092	0.0084	64,768		
Unmeasured	Predicted	Male	0.089	0.0052	109,916		
Unmeasured	Predicted	Female	0.108	0.0048	122,159		

Legend:
h^2: heritability
r_g: genetic correlation
SE: standard error

Table 4: LDSC GCTA estimated heritabilities (h^2) for observed and predicted AUDIT.

Set	Outcome	h^2 (SE)
AUDIT-Total		
Measured	Observed	0.0811 (0.006)
Unmeasured	Predicted	0.0727 (0.0036)
Measured	Predicted	0.08433 (0.0056)
AUDIT-Consumption		
Measured	Observed	0.0869 (0.0061)
Unmeasured	Predicted	0.0759 (0.0042)
Measured	Predicted	0.0816 (0.0056)
AUDIT-Problems		
Measured	Observed	0.0468 (0.0049)
Unmeasured	Predicted	0.0647 (0.0034)
Measured	Predicted	0.0769 (0.0056)

Table 5: LDSC estimated genetic correlations (r_g) between the observed and unobserved AUDIT in the measured and unmeasured sets.

Comparison	r_g (SE)	p-val
AUDIT-Total		
Observed vs. Predicted, Unmeasured	0.9746 (0.0354)	5.40E-167
Observed vs. Predicted, Measured	0.9191 (0.0181)	0
Predicted Measured vs. Predicted, Unmeasured	0.9746 (0.0333)	2.30E-188
AUDIT-Consumption		
Observed vs. Predicted, Unmeasured	0.8695 (0.0353)	3.30E-134
Observed vs. Predicted, Measured	0.858 (0.0222)	0
Predicted Measured vs. Predicted, Unmeasured	0.9712 (0.0349)	1.40E-170
AUDIT-Problems		
Observed vs. Predicted, Unmeasured	0.9126 (0.0583)	2.90E-55
Observed vs. Predicted, Measured	0.7915 (0.0387)	5.0E-93
Predicted Measured vs. Predicted, Unmeasured	0.9627 (0.0381)	5.30E-141

Figure 1: Flow logic of the MAGIC-LASSO procedure. The MAGIC-LASSO procedure begins with filtering, followed by clustering, then iterative Group-LASSO application until parsimony is achieved.

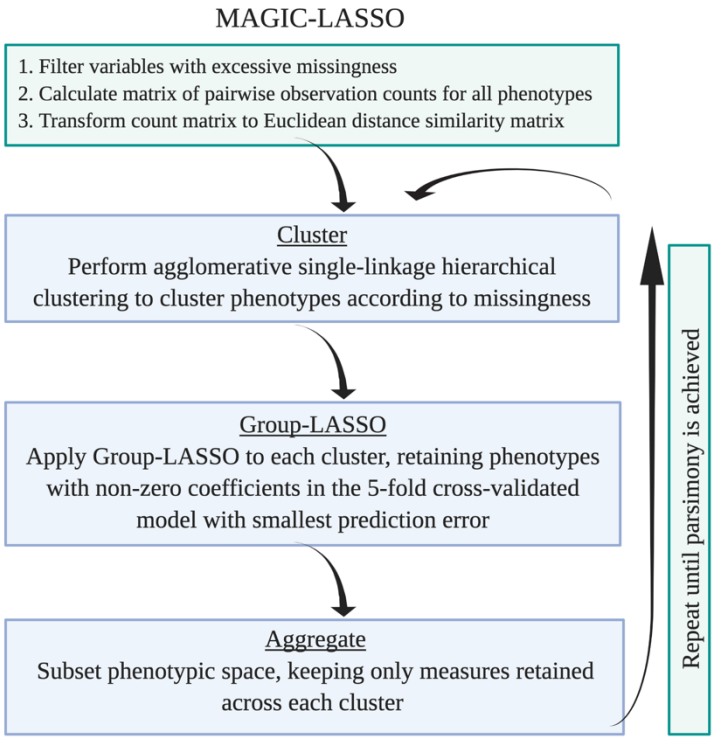


Figure 2: Conceptualization of how a dataset may be subdivided into a measured and unmeasured set. Where N represents the full sample size, $N_{Unmeasured}$ and $N_{Measured}$ represent the subsets of subjects on whom the outcome of interest is either missing or measured, respectively. Then the amount of overlap in observation may be quantified for each of p additional variables.

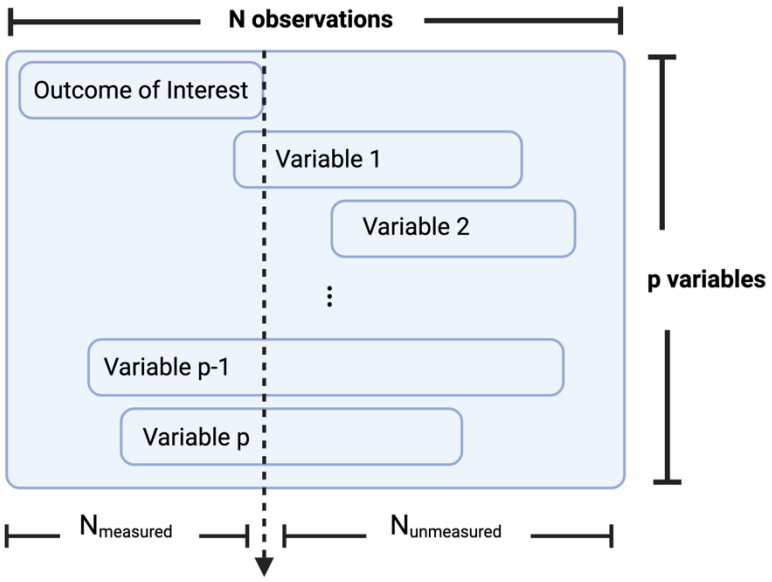


Figure 3: Illustration of clustering and Group-LASSO procedures within the MAGIC-LASSO. The hierarchical clustering steps divides the phenotypic space into non-overlapping subsets based on missingness. The Group-LASSO is applied to each cluster and variables retained by the algorithm in each set are aggregated.

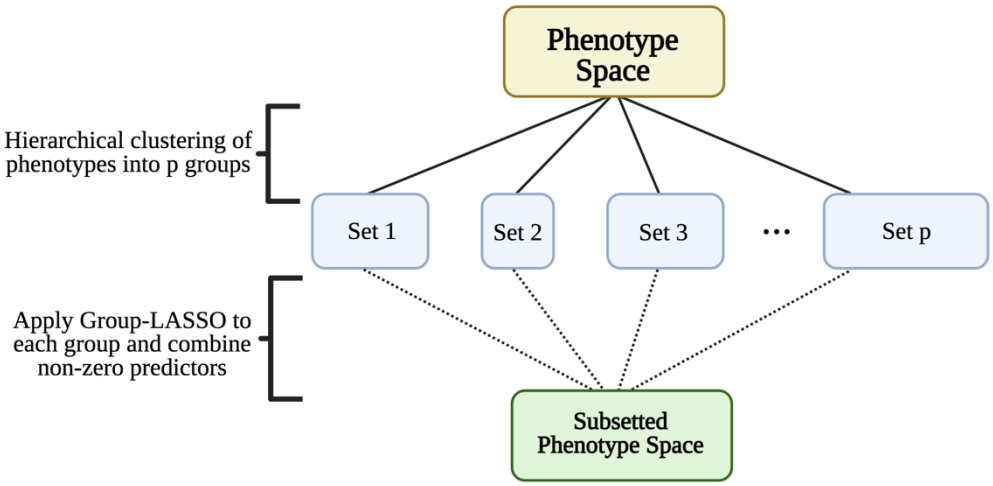


Figure 4: Flowchart of filtering in the UKB application set. The number of variables remaining after each filtering metric is applied is shown for the UKB data application example.

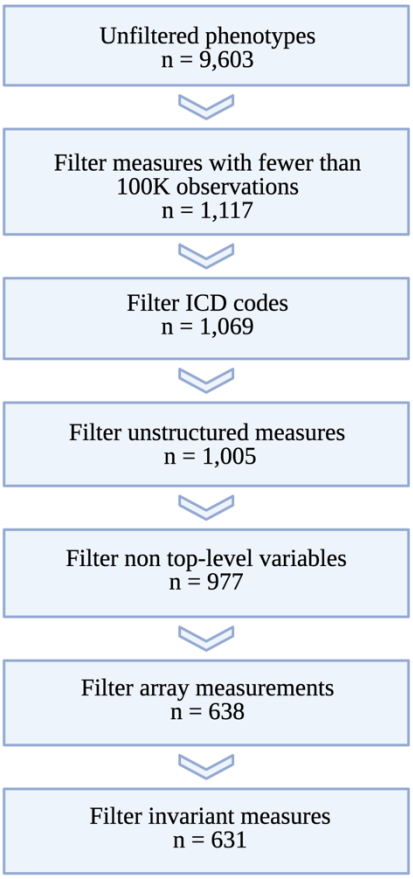
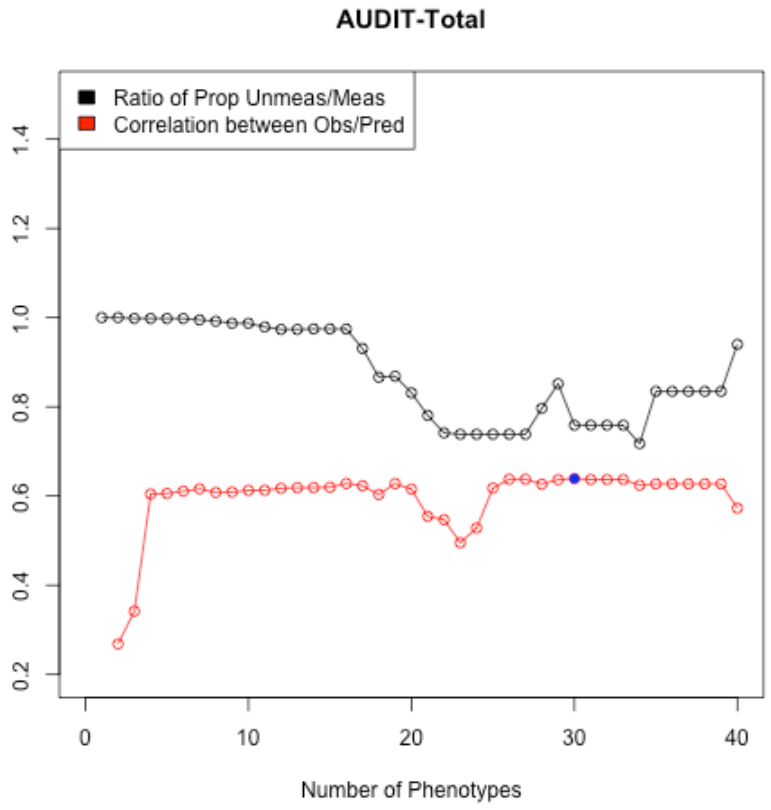
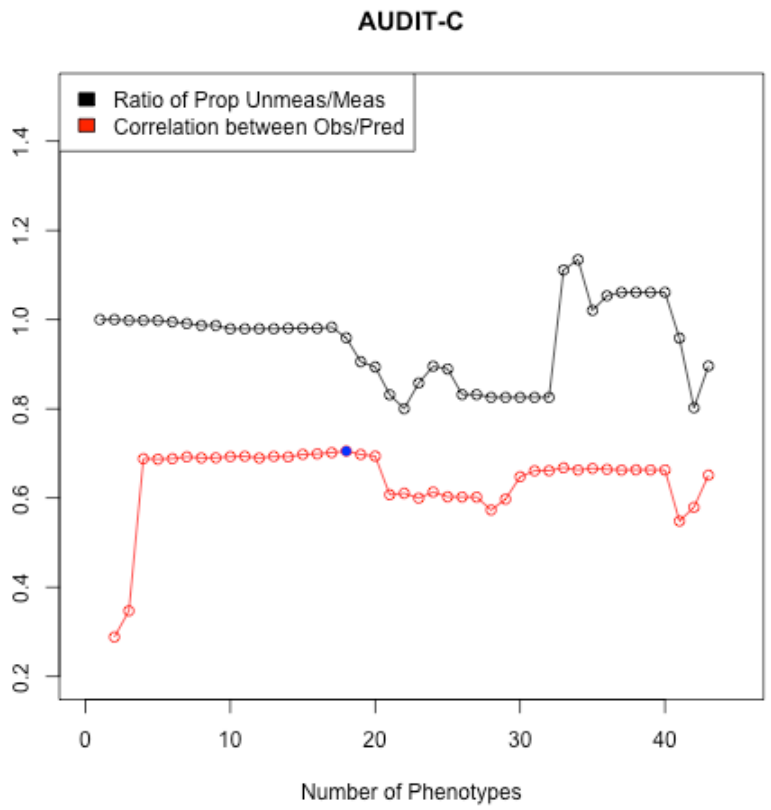


Figure 5: Stepwise model results. Phenotypic correlations and ratio of the proportion of total measured to unmeasured observations in the model set for (a) AUDIT-Total, (b) AUDIT-Consumption, and (c) AUDIT-Problems. Final selected model highlighted in blue.

(a) AUDIT-Total



(b) AUDIT-Consumption



(c) AUDIT-Problems

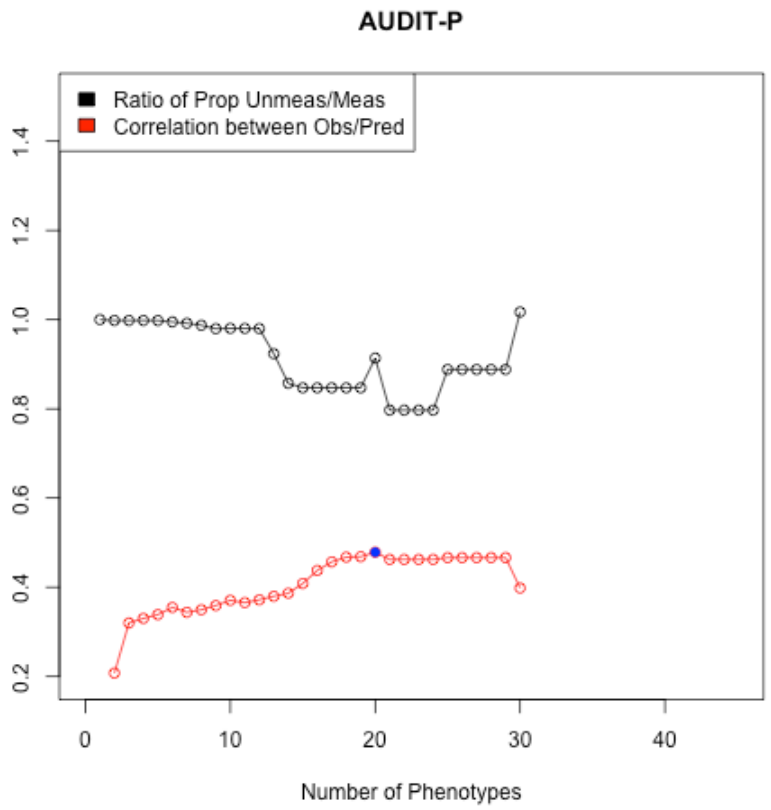
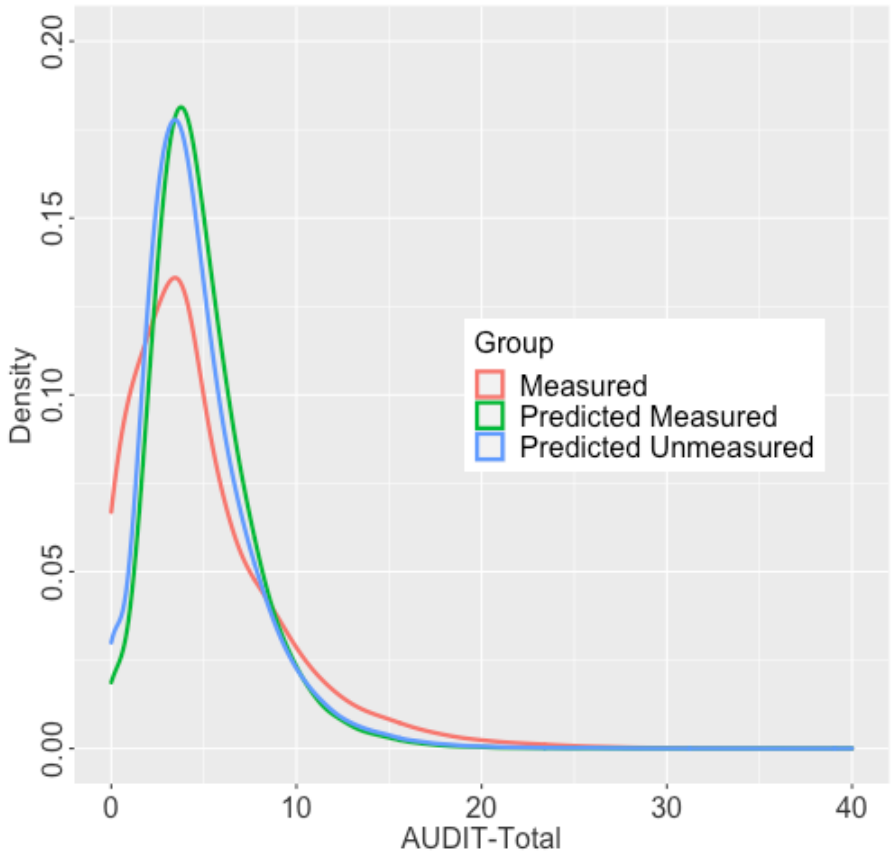
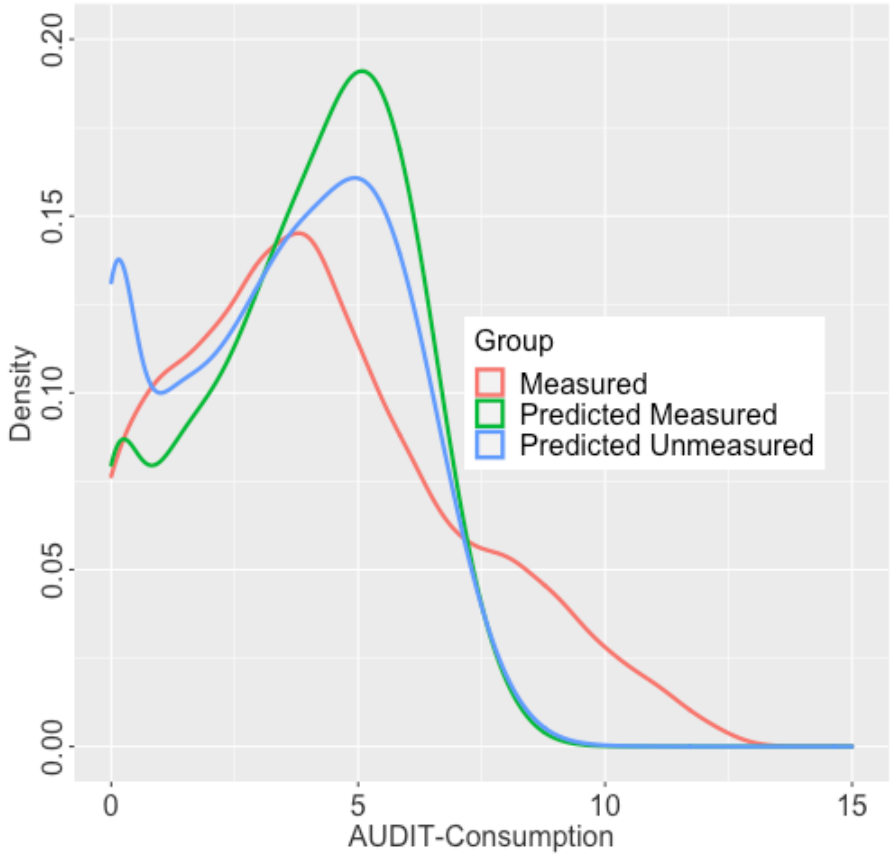


Figure 6: Density curves of the observed and predicted scores. Outcomes in the observed and predicted in the measured and unmeasured sets plotted for (a) AUDIT-Total, (b) AUDIT-C, and (c) AUDIT-C.

(a) AUDIT-Total



(b) AUDIT-Consumption



(c) AUDIT-Problems

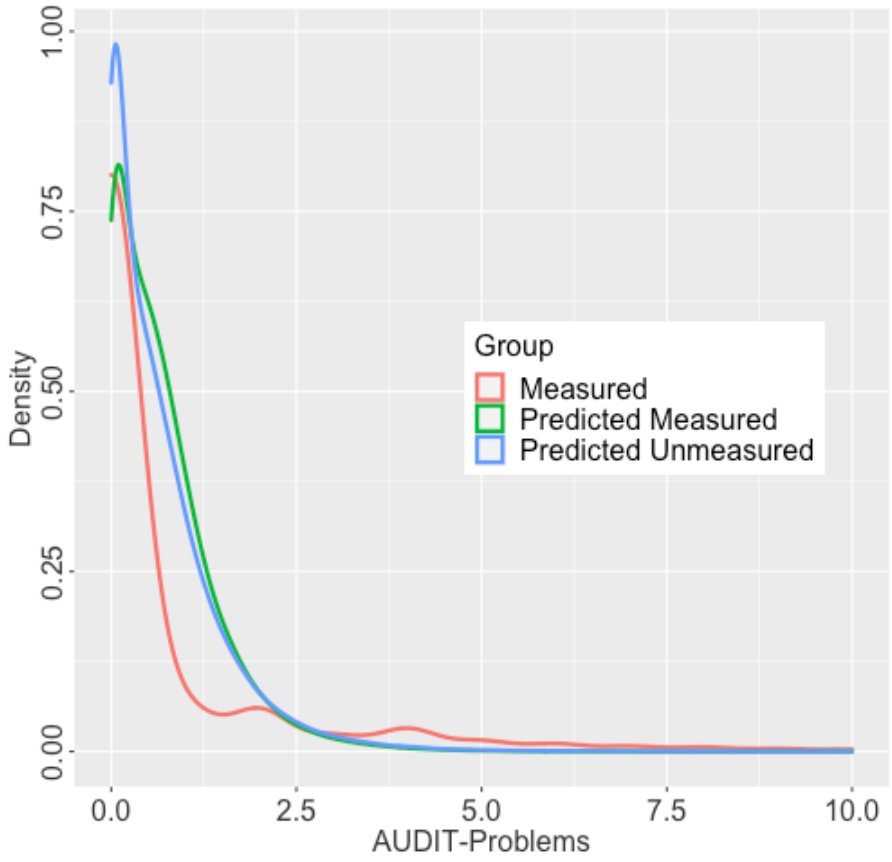
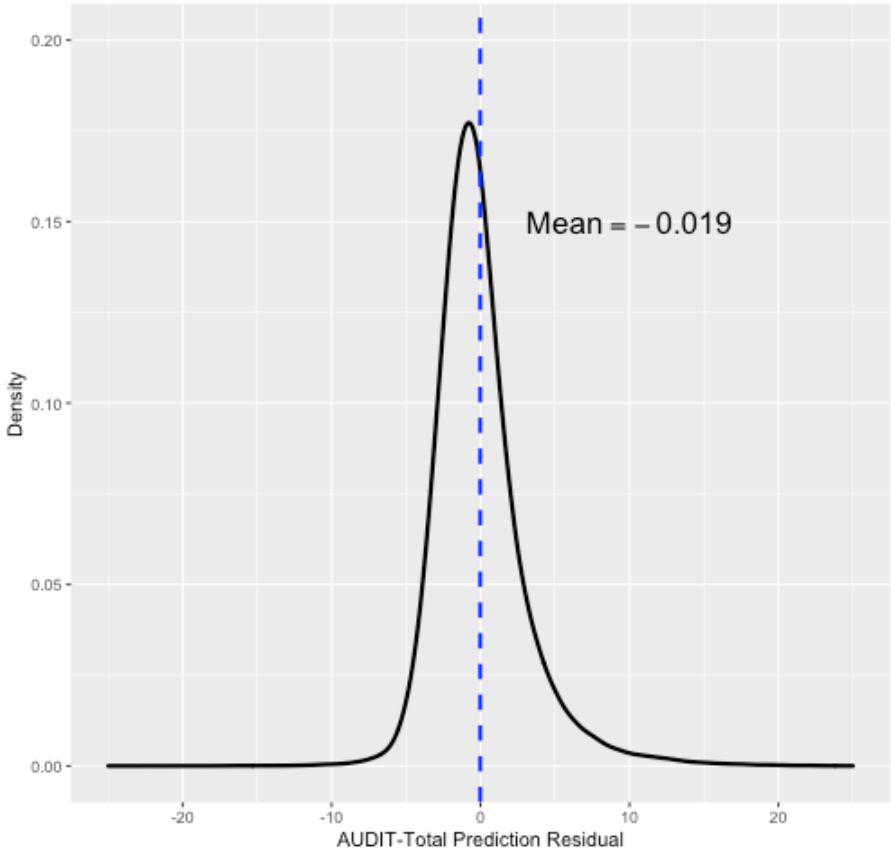
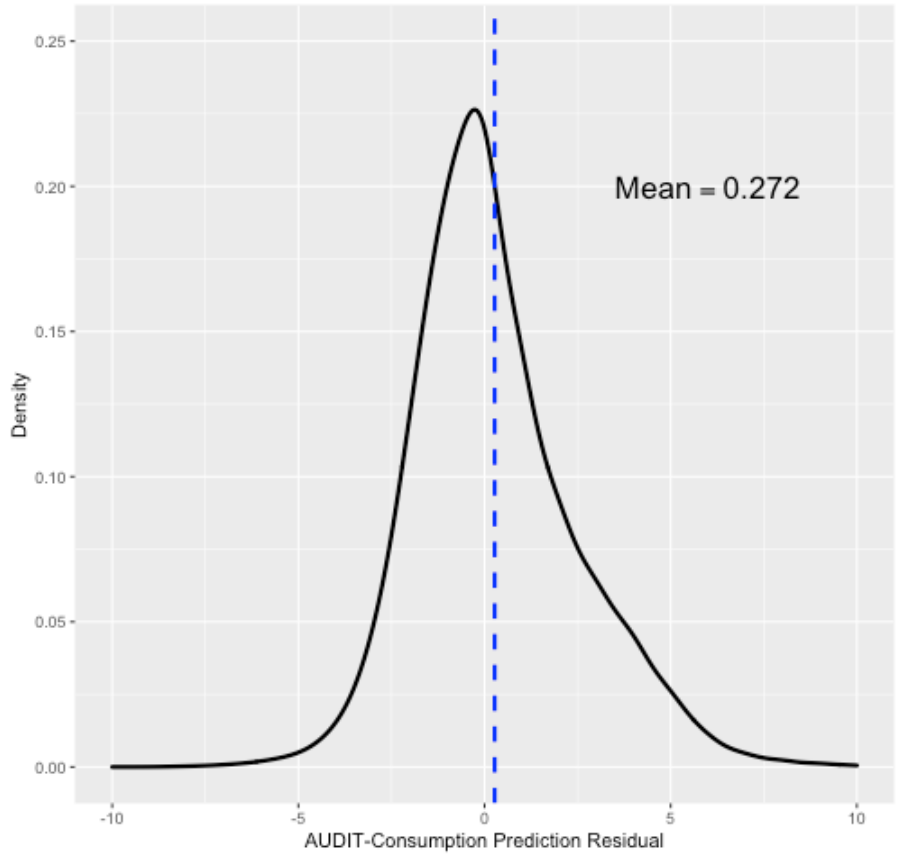


Figure 7: Residual densities for AUDIT prediction. Density curves with means noted showing the distribution of the prediction residuals for (a) AUDIT-Total, (b) AUDIT-Consumption, and (c) AUDIT-Problems.

(a) AUDIT-Total



(b) AUDIT-Consumption



(c) AUDIT-Problems

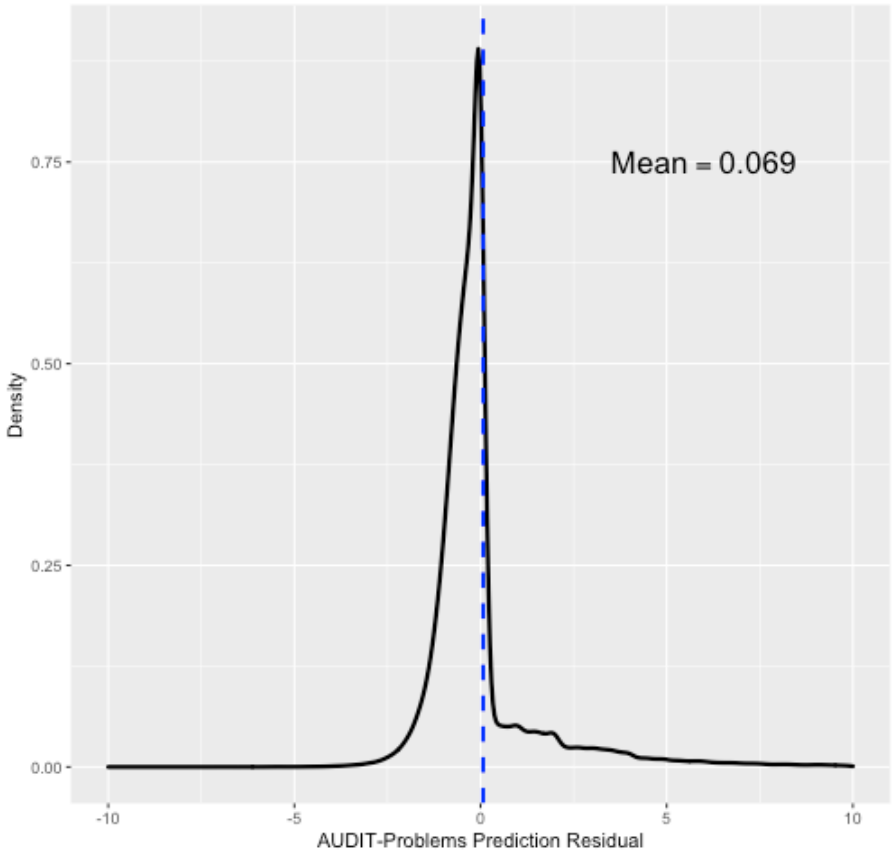


Figure 8: LDSC estimated heritabilities. SNP-based heritability estimates for the observed (green) and predicted in the measured (purple) and unmeasured (orange) sets for the AUDIT outcomes.

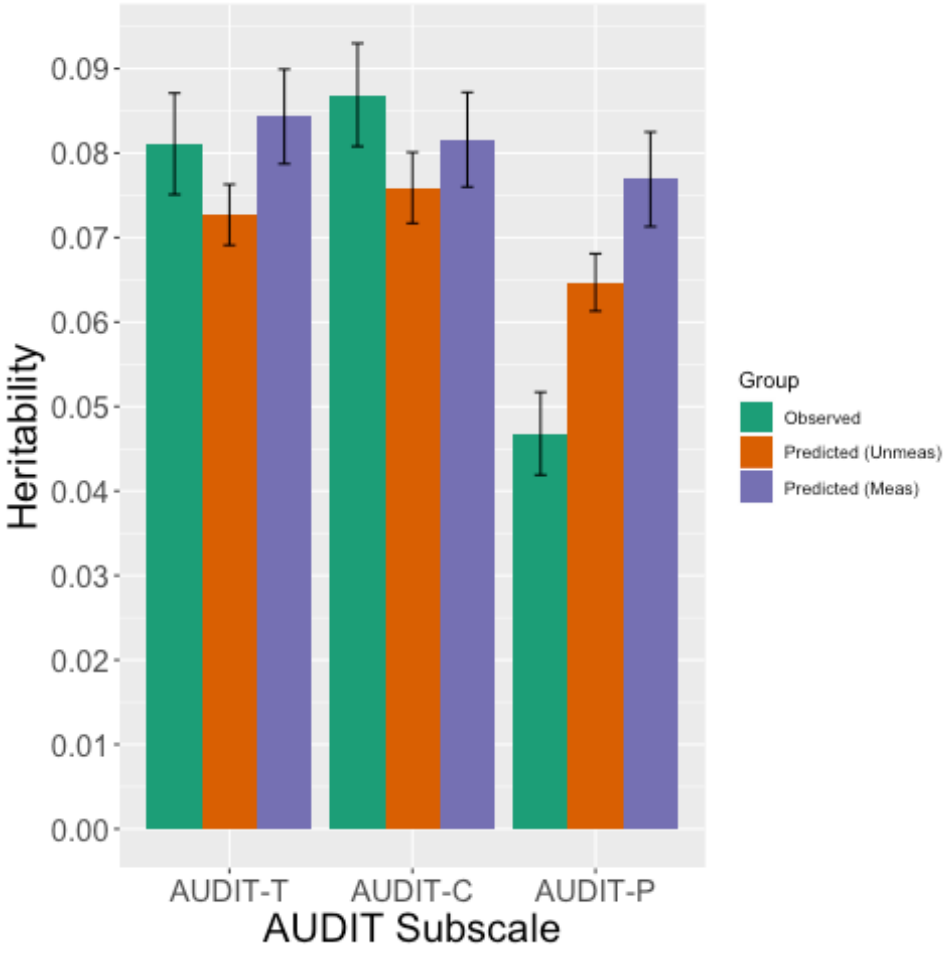
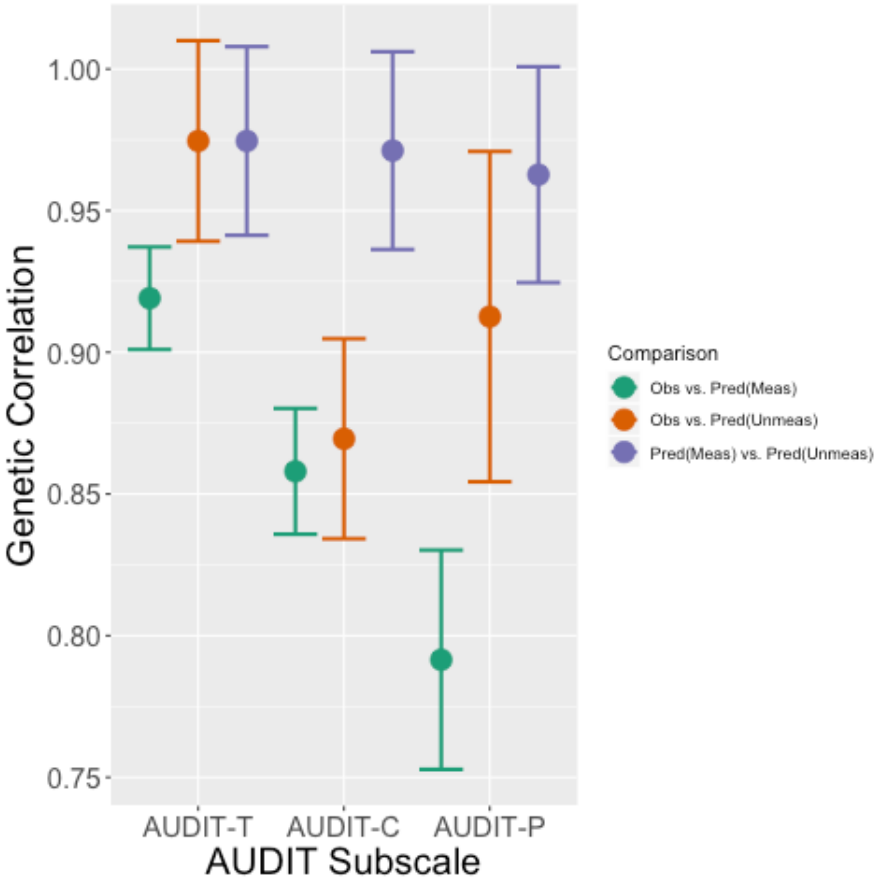



Figure 9: LDSC estimated genetic correlations. Genetic correlation estimated between the observed data and predicted scores in the measured sets (green,) the observed data and the predicted scores in the unmeasured sets (orange,) and the predicted scores in the measured and unmeasured sets (purple.)