White Paper 23-12

Estimating Complex Phenotype Prevalence Using Predictive Models

Authors:
Nicholas A. Furlotte
Aaron Kleinman
Robin Smith
David Hinds

Created: September 25th, 2015

Introduction

In this white paper, we describe an approach for creating and evaluating predictive models of human traits. At a high level, our goal is to systematically use demographic and genetic information to predict the likelihood that an individual exhibits a certain trait. We are primarily interested in approaches to predict categorical traits, in which one of *c* mutually exclusive eventualities occurs. In this case, the aim of a predictive model is to produce a vector of probabilities that expresses the likelihood that each of the *c* events will occur.

There are many machine learning methods that can be utilized to define a predictive model (logistic regression, support vector machines, decision trees, etc.). Typically, modeling is done on a case-by-case basis and the practitioner utilizes a mix of domain knowledge and statistical learning to test and ultimately define a final predictive model. Here, we define a standardized methodology to produce, test and validate predictive models for categorical human traits given a large amount of phenotypic and genetic data.

For a given trait, our computational pipeline selects relevant predictors, builds a predictive model and then produces a set of evaluation metrics. Our pipeline sets a basic standard for high-throughput predictive modeling of human traits that we will later build on when interrogating more complicated phenotypes and using more sophisticated models.

Methods

Computational pipeline for defining predictive models of human traits

The structure of our computational pipeline is illustrated in Figure 1. We begin by selecting a trait of interest from the collection of phenotypic data at 23andMe. This choice is informed by a literature search, domain knowledge, presumed customer interest and internal GWAS results. Next, given the full cohort of participants with relevant data for our selected trait, we filter out related individuals and split the cohort into a training and validation set. We then run a GWAS for feature selection, and choose strongly associated SNPs as predictors. Next, we test and evaluate predictive models using these features. This step may also include custom refinements, such as adding well-known SNPs not identified in the GWAS. Lastly, we test models over our training set, define a "final model" and compute performance statistics in our independent validation set.

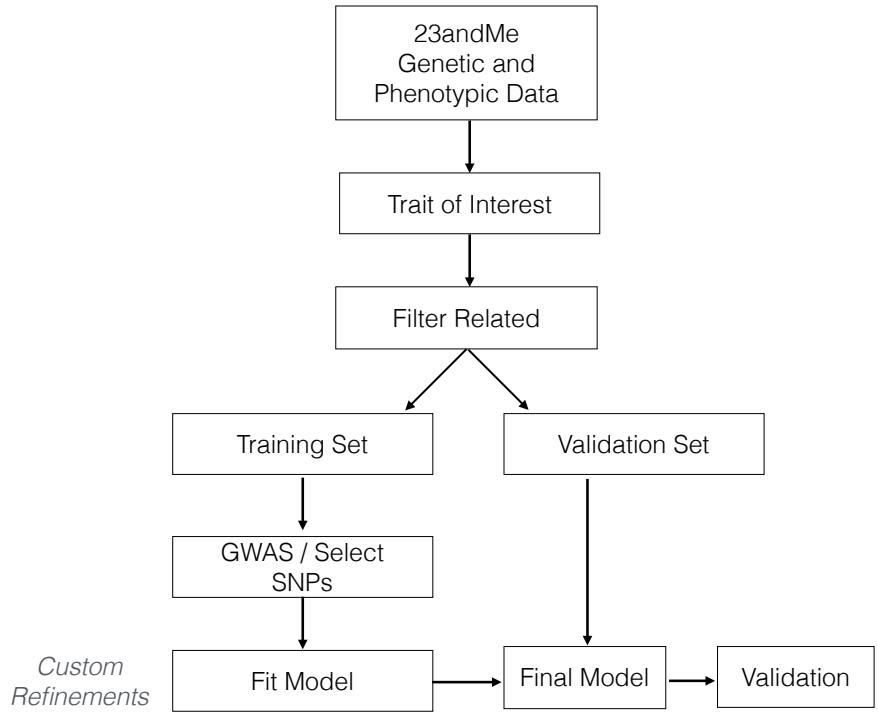

Figure 1: Overall structure of pipeline.

Predicting categorical phenotypes

One of the primary goals of statistical analysis is to make predictions about unknown and future events, and one of the most reasonable ways to present these predictions is as a probability distribution over the possible events (Dawid 1984). The basic idea is that given an event E with c mutually exclusive possible outcomes $\{E_i\}$, and a set of predictors X, we would like to predict the probability of each event given the observed predictors X or $P(E_i|\, X)$. For our purposes, events are phenotypic labels and predictors are both demographic information and genetic variant information. For example, when considering the phenotype hair color, we would like to predict the probability of the events blond, brown, black or red hair and might use age, sex and genetic variants in the OCA2 gene as predictors.

There are many ways to produce predictions of the form $P(E_i|\, X)$. One approach is to use logistic regression for binomial outcomes and multinomial logistic regression for the more general case of multinomial outcomes. Support vector machines, K-nearest neighbors, decision trees and neural nets are all machine learning methods that could also be used to train probabilistic forecasting models.

Selecting features

In version 1 of our prediction pipeline, we assume that our possible pool of features includes age, sex and all genotyped SNPs. We determine if age and/or sex should be included in the predictive model on a case-by-case basis. For example, sex is not used as a predictor for a trait that only applies to males such as "male pattern

baldness", but would be used in a trait that applies to both males and females (eg. hair color).

SNPs are selected through a straightforward procedure. First, we run a standard genome-wide association analysis on the trait of interest over our training set. Next, we identify a set of peak SNPs. Peaks are identified by selecting a single SNP to represent 500kb regions that contain at least one association with p-value less than or equal to 5e-6. With a list of peak SNPs and associated p-values, we select the set of peak SNPs with p-value less than or equal to 5e-8 as our final set of predictors. This list may be refined further through curation of the literature.

These SNPs, along with age and sex, become the features input to our machine learning algorithm.

Defining the model

In internal research, we tested a variety of machine learning approaches for a number of different traits. Generally, we found that logistic regression; multinomial logistic regression and ordinal regression performed well and often substantially outperformed other standard machine learning algorithms, such as decision trees. In addition, logistic regression is the standard technique used in GWAS analysis for binary traits. Given these considerations and the overall simplicity of regression-based methods, we chose to use logistic regression for binary traits, ordinal regression for categorical traits that have a natural ordering and multinomial logistic regression for categorical traits without an inherent ordering.

The first step in these methods is to define a training set and validation set. To do this we first define the full cohort for a given trait by identifying all individuals with non-missing data. Next, we remove individuals from the cohort until there are no two individuals that share more than 700 centimorgans of computationally determined identity by descent (IBD). This is done to ensure that close relatives do not bias the estimated parameters of the model. Finally, we randomly select 90% of the individuals to compose the training set and let the remaining 10% comprise the validation set.

As mentioned in the feature selection section, a GWAS is run on the training sample and SNPs are selected from this GWAS. We use a standard implementation of logistic regression and a custom version of ordinal regression assuming the proportional odds model. Our multinomial regression implementation uses an all-vs-one approach where each all-vs-one logistic model is fit and the final probabilities for each class are combined using the softmax function.

Each method can be used to obtain a continuous score for each individual. In the case of a logistic model, the continuous score is the probability of the event encoded as case. For ordinal regression it is the weighted sum of the predictors minus the label specific intercept and for multinomial logistic models each phenotypic label is associated with a continuous probability. Using the appropriate score, we place

training set individuals into 20 bins each representing 5% of the total training set, such that the first bin represents individuals with the 5% lowest scores and the last bin represents the individuals with the highest 5% of scores and so on. Using this binning logic, we calculate the frequency of each phenotype label in each bin. These serve as our final predictions. In other words, we can use the model to calculate a score and determine what bin the individual falls into and then their prediction is simply the frequency of the phenotypic labels in the appropriate bin in the training set. In practice, we find that this method leads to better calibration and is generally easier to interpret and evaluate since there are only 20 potential outcomes for the customer.

Before defining the final model and evaluating in the validation set, there is an opportunity to make custom refinements. One typical custom refinement is to add SNPs that are well known to affect the condition, but that might not have been identified as part of the GWAS (for instance, SNPs from the same locus that are not in linkage disequilibrium).

Evaluating the model

Calibration

Calibration, also called reliability, refers to the consistency of the model or its ability to produce probability estimates that are concordant with observed frequencies (Gneiting 2007). For example, if we take everyone who is predicted to have 20% chance of having a given condition, then we expect that 20% of those people actually report having the condition. If that is the case across the full probability distribution, then the model is fully calibrated.

For a case-control phenotype, where each individual can be placed into one of two categories, calibration is evaluated visually using a calibration plot. In a calibration plot, the x-axis represents the expected proportion of cases and the y-axis represents the observed proportion of cases. The statistical significance of the level of calibration is often evaluated through the Hosmer-Lemeshow statistical test. We found that this statistic coupled with a visual inspection works well to evaluate calibration.

Area under the curve (AUC) and multinomial AUC (mAUC)

Area under the receiver operating characteristic curve (AUC) is often used as a way to evaluate the discriminatory power of a predictive model. AUC has a simple interpretation. Consider a model score that can be interpreted as the probability of having a given disease. AUC is then defined as the probability that a randomly sampled control has a lower probability of disease when compared with a randomly sample case. This measure has range 0 to 1. A value of 0.50 is equivalent to random, while a value of 1 means the model discriminates perfectly between cases and controls.

Because AUC is not appropriate for categorical models with more than two outcomes, the literature contains a number of suggested generalizations of AUC to

deal with this case. We use measure proposed by Hand and Till (2001) which we call mAUC and which is defined as follows. Assume our trait has c classes: 1, 2, 3,..., *c*. Let A_{ij} represent the probability that a randomly drawn member of class *i* will have a higher probability of belonging to class *i* than a randomly drawn member of class *j*. This is essentially the AUC measure when class *i* represents cases and class *j* represents controls. Note that A_{ij} does not necessarily equal to A_{ji}. Define $D_{ij} = (A_{ij} + A_{ji}) / 2$ and then define mAUC as $\frac{1}{\binom{c}{2}} \sum_{i,j} D_{ij}$

The mAUC measure is simply the average over the average pairwise AUCs for each pair of classes, and indeed it equals the standard AUC when *c*=2. It can be evaluated in a similar fashion to AUC: mAUC equal to 0.50 is equivalent to random and closer to one is better.

Results

The prediction pipeline takes as input a phenotype of interest, generates a training and validation set, runs a GWAS in the training set from which it selects significant SNPs to use as predictors, fits the appropriate model (logistic, multinomial logistic or ordinal) and generates a set of outputs that are used to evaluate the model. The outputs consist of the following:

1. **Demographic data** for the training and validation sets.
2. **Distribution of model scores**. This is depicted as a histogram, showing the fraction of individuals (density) that have each model score.
3. **ROC curves and AUC**. For multinomial models, we produce an ROC curve for each phenotype label and compute mAUC. We use bootstrap to estimate confidence intervals.
4. **Training set frequency**. We use a score to place individuals in the training set into vigintiles and then compute the frequency of observed labels (e.g. all cases, cases with light hair, etc.) in these bins. The score for logistic models is the predicted probability of the case label and for ordinal models it is the continuous threshold score. For multinomial models, the table is generated with a specified class probability.
5. **Calibration plots**. For multinomial models, we produce calibration plots for each label. Calibration plots are created by plotting the expected frequency of the label on the x-axis against the observed frequency of the label on the y-axis in each vigintile. The expected frequency of the label is simply the average probability reported for an individual in a given vigintile.

These outputs are displayed in templated HTML format so that each model can be systematically evaluated. Below we will show examples of the structure of these outputs for a particular phenotype: attached earlobes.

Case Study: Attached Earlobes

Phenotype Demographics

Phenotype	Label	Total	(0,30]	(30,45]	(45,60]	(60,Inf]	Female	Male
earlobes	Unattached	41590	5594	11649	11020	13327	20530	21060
	Attached	14137	1845	4153	3747	4392	7640	6497

Table 1: Demographic information for the "attached earlobe" phenotype. For each phenotype we produce a table that shows the number of individuals that have reported each phenotypic label and stratify this by age range and sex.

Distribution of Model Scores

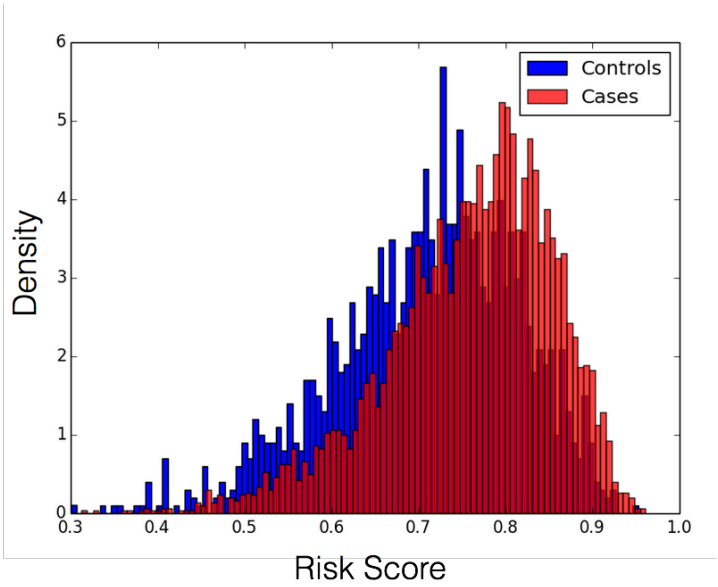

Figure 1: Distribution of model scores for the "attached earlobe" phenotype. Earlobe attachment is a case-control phenotype: the case label represents individuals with unattached earlobes. The predictive model used is logistic regression and the model score is the probability of being a case (having unattached earlobes). We see that the case distribution is shifted to the right with respect to the control distribution indicating that cases on average have a higher model scores when compared to controls.

ROC Curve and AUC

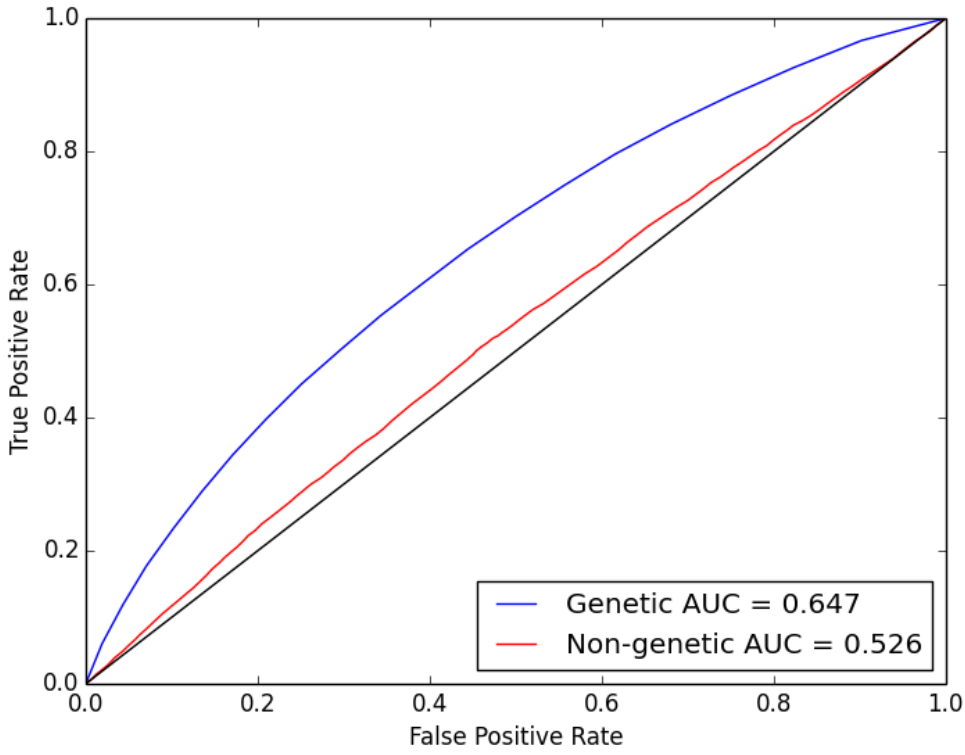

Figure 2: ROC curve for attached earlobe phenotype. We display two ROC curves and the corresponding area under the curve estimates (AUC): one for the model including both demographic and genetic information (Genetic) and one for the model including only demographic information (Non-genetic). We expect that the genetic model will have a higher AUC when compared with the non-genetic model.

Calibration

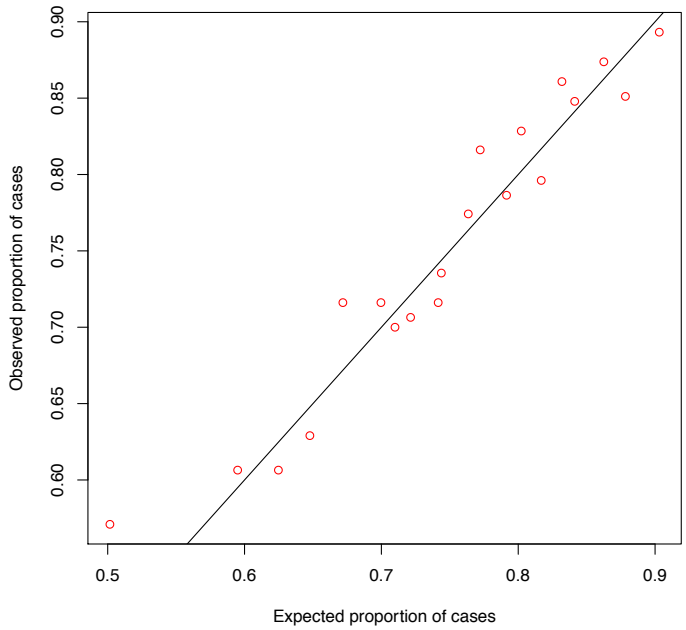

Figure 3: Calibration plots for attached earlobe phenotype. Calibration assesses the degree to which the estimated probability of being a case matches the observed frequency of cases. We expect deviations from the diagonal line and see that for some vigintiles their "risk" is over or under estimated. Part of the evaluation process is to assess and qualitatively evaluate the magnitude of these deviations. In this case, a Hosmer-Lemeshow test fails to reject calibration (p = 0.15).

Training Set Frequency

Vigintile	min score	max score	attached	unattached
1	0.5	0.6	50%	50%
2	0.6	0.63	40%	60%
3	0.63	0.65	37%	63%
4	0.65	0.67	35%	65%
5	0.67	0.7	33%	67%
6	0.7	0.71	30%	70%
7	0.71	0.72	29%	71%
8	0.72	0.74	28%	72%
9	0.74	0.74	26%	74%
10	0.74	0.76	26%	74%
11	0.76	0.77	24%	76%
12	0.77	0.79	23%	77%
13	0.79	0.8	21%	79%
14	0.8	0.82	20%	80%
15	0.82	0.83	18%	82%
16	0.83	0.84	17%	83%

17	0.84	0.86	16%	84%
18	0.86	0.88	14%	86%
19	0.88	0.9	12%	88%
20	0.9	1.0	10%	90%
Overall	**0.5**	**0.9**	**25%**	**75%**

Table 2: Training set frequency table for attached earlobe phenotype. This table shows the frequency of each label (attached or unattached) in each vigintile (5% of the training set). The last row shows the overall frequency of the labels (population prevalence). We see that around the 10th and 11th vigintile, individuals have average probability (the population prevalence), while in the extremes (1st and 20th) they have probabilities that deviate significantly from the population prevalence.

Study Participants

Study participants were 23andMe customers who had been genotyped as part of the 23andMe PGS®, consented to research, and voluntarily responded to web-based surveys. The study protocol and consent form were approved by the external Association for the Accreditation of Human Research Protection Programs-accredited Institutional Review Board, Ethical and Independent Review Services.

All study participants were required to have > 97% European ancestry, as determined by analysis of local ancestry [12]. The reference population data for ancestry analysis were derived from public datasets (the Human Genome Diversity Project, HapMap, and 1000 Genomes) and from 23andMe customers who have reported having four grandparents from the same country. At present, the database has the highest power to detect associations in cohorts of European ancestry.

We also required that participants were unrelated (sharing a smaller percent of the genome than the minimum expected for first cousins), as determined by a segmental identity-by-descent estimation algorithm [13].

Sample collection, DNA extraction, and genotyping

Participants provided their saliva using a specialized collection device, and mailed their samples to a CLIA-certified laboratory contracted by 23andMe. DNA extraction and 23andMe BeadChip genotyping were performed by the contracted laboratory [1,6]. All samples were genotyped on one of four versions of 23andMe BeadChips (produced by Illumina) that probe over half a million single nucleotide polymorphisms (SNPs) distributed across the entire human genome.

Acknowledgements

We thank the customers of 23andMe for answering surveys and participating in this research. We also thank all the employees of 23andMe, who together have made this research possible.

References

Gneiting, Tilmann, Fadoua Balabdaoui, and Adrian E. Raftery. "Probabilistic forecasts, calibration and sharpness." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69.2 (2007): 243-268.

Dawid, A. Philip. "Present position and potential developments: Some personal views: Statistical theory: The prequential approach." Journal of the Royal Statistical Society. Series A (General) (1984): 278-292.

Bishop, Christopher M. Pattern recognition and machine learning. Vol. 4. No. 4. New York: springer, 2006.

Lobo, Jorge M., Alberto Jiménez-Valverde, and Raimundo Real. "AUC: a misleading measure of the performance of predictive distribution models."Global ecology and Biogeography 17.2 (2008): 145-151.

Ranjan, Roopesh. Combining and evaluating probabilistic forecasts. Diss. University of Washington, 2009.

Bickel, J. Eric. "Some comparisons among quadratic, spherical, and logarithmic scoring rules." Decision Analysis 4.2 (2007): 49-65.