A REWARD SYSTEM POLYGENIC RISK SCORE FOR PREDICTING OBESITY AND SUBSTANCE ADDICTION

By

Kristen M. Stevens

Bachelor of Arts, University of California, Berkeley, 2010

A DISSERTATION

Presented to

the Department of Medical Informatics & Clinical Epidemiology

and the Oregon Health & Science University

School of Medicine

in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

June 2020

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the PhD dissertation of

Kristen M. Stevens

has been approved

_____

Shannon K. McWeeney, Ph.D., Mentor

_____

Guanming Wu, Ph.D., Chair

_____

Joyanna Hansen, Ph.D., R.D., Member

_____

Daniel L. Marks, M.D., Ph.D., Member

_____

Suzanne H. Mitchell, Ph.D., Member

It was when I said,

"Words are not forms of a single word.

In the sum of the parts, there are only the parts.

The world must be measured by eye."

— Wallace Stevens, "On the Road Home"

§

Everything about her spoke of alternatives and possibilities that

if considered too deeply would wreak havoc with

the neat plan I had laid out for my life.

— Tsitsi Dangarembga, *Nervous Conditions*

For all of my teachers and all of my students—

we will go forth together.

TABLE OF CONTENTS

Acknowledgements

I would like to offer my deepest acknowledgements to my mentor, Dr. Shannon McWeeney, for her unending commitment to teaching me the beauty of the scientific method, for preserving (and at times rekindling) the awe and wonder that first led me to a career in science, for practicing the patience and rigor that I will bring to every scientific endeavor in my future, and lastly for her light-hearted guidance and friendship. Thank you for nudging me along this journey; there is no way I would be the scientist (or teacher) I am today without you.

Next, I would like to thank the members of my dissertation advisory committee, who made working on this group project together especially meaningful. To my committee chair, Dr. Guanming Wu, for always holding me to the highest standards and opening my eyes to all the creative ways network methods can solve seemingly intractable problems in biology and medicine. To Dr. Joyanna Hansen, for believing in me early on during my time as a graduate student, offering me the opportunity to teach and share my excitement in using genomics to understand more about human nutrition, and always asking the clarifying questions necessary to make this work approachable to a wide scientific audience. And to Dr. Dan Marks, for remembering what it was like to be an idealistic MD-PhD student who wants to change the world for the better, for his willingness to trust my judgment even when he probably shouldn't have, and his commitment to making this work the best it could possibly be.

I would like to acknowledge Dr. Aurora Blucher for her ongoing mentorship; it is very unlikely I would have ever come to embrace the utility of mathematical modeling for solving real-world scientific problems without your support. I would like to acknowledge the faculty and post-doctoral scholars that offered key insights into this work including Dr. Lucia Carbone, Dr. Eilis Boudreau, Dr. Suzanne Mitchell, Dr. Michael Mooney, Dr. Jessica Minnier, Dr. Beth Wilmot, Dr. Reid Thompson, Dr. Dian Chase, and Dr. Eric Feczko. I would also like to

acknowledge the faculty with direct involvement in my early scientific training during graduate school, Dr. David Koeller and Dr. Kent Thornburg, whose support and encouragement were crucial to embarking on the work presented here. Furthermore, I would like to acknowledge Dr. William Hersh and Dr. David Jacoby, for their unwavering support of my scientific education and for building two leading research education programs without which this work would have been impossible.

I cannot imagine completing this work without the camaraderie and inspiration of all the graduate students, post-doctoral scholars, and clinical fellows who were part of the Department of Medical Informatics & Clinical Epidemiology (DMICE) over the years. I especially would like to thank my closest fellow graduate students: Josh Burkhart, Julian Egger, Eric Leung, Rose Goueth, and Ben Cordier. I would like to acknowledge the DMICE faculty and staff for providing such a rich scientific community and support throughout my time as a graduate student in the department; I would especially like to thank Diane Doctor, Virginia Lankes, Lynne Schwabe, Andrea Ilg, and Lauren Ludwig.

I also cannot imagine completing this work without all of my fellow students in the MD-PhD Training Program. I would especially like to acknowledge Elizabeth Swanson for her ongoing peer mentorship and friendship starting from day one of medical school in anatomy lab; you have a perseverance that continues to inspire me to this day. I would also especially like to acknowledge the program coordinators, Johanna Colgrove and Alexis Young, for their help in navigating this nearly decade-long path as an MD-PhD student.

Last, but certainly not least, I would like to thank my closest friends and family. To Nathalie Javidi-Sharifi and Johannes Elferich, words will never come close to expressing the gratitude I feel towards the two of you; you have fed my mind and belly to the brink over the

past seven years, and while I am sad that our time together is coming to an end for now, I am looking forward to watching from afar as your next chapter unfolds. To Scott and Laurel Hoffmann, Burcu Gurun-Demir, Rachel and Gary Sivek, Addie Cuneo, Jon Buckalew, Maansi Shah, Janet Slesinski, Morgan Maschmeier, Catherine Welch, Lea Juzek, Allissia Gilmartin, Erika Freeman, and Danika Robison, your friendship and encouragement over all these years have made this process that much easier.

To my mother- and father-in-law, Hope and Michael, thank you for all of the support and kind words over the years; you have taught me more than you can imagine. To my Aunt Carol, and my cousins Eric and David, your unwavering support and confidence in me made me believe that my success as a scholar was inevitable so long as I just stuck with it. To my "Ant" Jan, whose commitment to a life of service and caring for others has kept me grounded throughout this time devoted to academics. To my sister, Sydney, and my brother, Casey, thank you for keeping the fun and humor in my life alive despite my relentless (and mostly unsuccessful) attempts to become "serious." To my Mommom, who reminds me: "No patience, no friends," and that it is our relationships that are the most important part of life. And of course, to my mom and dad, Kimberly and James, who have made immeasurable sacrifices to get me where I am today, always ready to wipe away my tears of frustration and a hug to keep me going.

To my son, Heath, you have been one of the greatest adventures of my life, and I am so looking forward to sharing all of this work with you someday. And to my husband and best friend, Reid, your steadfast belief that we are accomplishing great things together on this journey called life keeps me dreaming and aspiring for more.

Abstract

A Reward System Polygenic Risk Score for Predicting Obesity and Substance Addiction

Kristen M. Stevens

Oregon Health & Science University, June 2020

Genome-wide polygenic risk scores (PRS) can now predict complex genetic disease risk with nearly the same ability as tests for monogenic diseases. Despite this, there is no clear consensus how to incorporate PRS with other known modifiable and non-modifiable risk factors at the point of care. This challenge is further complicated by the fact that the most promising diseases for early PRS adoption (e.g., coronary artery disease, type 2 diabetes, and breast cancer) share many of the same modifiable risk factors—specifically, diet-induced obesity and drug use. Interestingly, the cause of these modifiable risk factors is at least partially genetic in most people. And while evidence for a common biological basis underlying nutrient intake and drug use in humans is growing, current clinical risk prediction models for complex genetic diseases have not incorporated any of this shared biology.

Here we develop a reward system aggregate phenotype score from daily quantitative, self-reported nutrient intake and drug use in 57K people of white British ancestry. We conduct a genome-wide association study to identify reward system-associated loci, and then construct a reward system PRS from these loci to predict individuals with obesity or substance addiction. We also include pertinent environmental risk factors that could be reasonably ascertained at a clinic visit (i.e., home location's neighborhood poverty rate and history of child abuse) in our prediction models. While our reward system PRS was not able to improve prediction in a separate test set of 12K people as compared to an obesity PRS or a substance addiction PRS, this research provides the necessary foundation for future preventative and therapeutic precision medicine efforts in the area of obesity and substance-related addictive disorders.

INTRODUCTION

Genome-wide polygenic risk scores (PRS) can now predict complex genetic disease risk with nearly the same ability as tests for monogenic diseases.[1] However, there are substantial barriers to implementing such tests in the clinic, most notably how to incorporate PRS at the point of care with other known modifiable and non-modifiable risk factors.[2-7] Some of the most promising diseases for early PRS adoption—such as coronary artery disease, type 2 diabetes, and breast cancer—share a common set of robust modifiable risk factors that includes diet-induced obesity and drug use.[8,9] Yet the cause of both obesity and substance use disorders is at least partially genetic in most people.[10-15] Moreover, intriguing evidence for shared underlying biology between nutrient intake and drug use in humans continues to accumulate.[16,17] Despite these advancements, clinical risk prediction models for complex genetic diseases have not incorporated any of this shared biology.[18-20]

At present, greater than 13 percent of the world's population is obese, while global deaths attributed to alcohol and tobacco are 5 and 12 percent, respectively.[21-23] According to the U.S. National Comorbidity Survey, however, obesity and a current substance use disorder do not co-occur in individuals more frequently than is expected by chance, given their respective prevalence and incidence rates in a general population.[24] Regardless, theoretical models support a common biological basis for obesity and substance use disorders in humans, whereby drugs and certain nutrients found in food compete for overlapping reward mechanisms, thus reducing the probability of coincidence of both conditions in a single person.[25-27] Recent empirical evidence from genome-wide association studies (GWAS) lends support to this theory.[28-30] Specifically, genetic variants in fibroblast growth factor 2 and β-klotho—which in humans are both associated with macronutrient intake and alcohol consumption—function in non-human

primate models as a ligand-receptor pair and nutrient sensor that controls overall food intake.[31] Meanwhile, positron emission tomography studies of obesity, binge eating disorder, alcohol use disorder, and cocaine use disorder all report loss of μ-opioid and dopamine D2 receptors in cortical, subcortical, and striatal brain regions.[32,33] Further evidence from lesion studies report associations between damage to specific regions of the frontotemporal lobes, weight change, and disruption of smoking addiction.[34,35] However, before recent monumental achievements in large-scale biomedical data collection, storage, sharing, and analysis, a population-level investigation of the biology common to obesity and substance use disorders was, from a practical standpoint, infeasible.[36-38]

While there is a substantial genetic component to both obesity and substance use disorders,[39-43] accounting for variation due to an individual's environment and lifestyle makes pursuing an observational study (similar to the one presented here) challenging but not impossible.[44-48] Thanks to previous work by researchers in this area, a small number of early-life environmental exposures have emerged as leading risk factors for both diet-induced obesity and substance use disorders—namely, living in a neighborhood with a high rate of poverty (including unreliable access to food) and a history of physical or sexual abuse.[49-51] In addition, lifestyle measures (i.e., modifiable risk factors) are now routinely self-reported by the participants of large population cohort studies through web-based surveys, which provide a more nuanced picture of the behaviors implicated in the etiology of obesity and substance use disorders.[37,52,53] Recognition that these behaviors, including fat and sugar intake, caffeine use, alcohol consumption, and cigarette smoking, exist on a continuum with a non-zero genetic contribution has led to further insights into disease mechanisms.[28,29,54-59] Of particular interest is the combination of genetic variants found in these studies reflecting both substance-specific

metabolism and reward-motivated behavior. These findings allude to an unexploited source of genetic variation with the potential to improve disease prediction, and encouraged us to perform a systematic evaluation of the biology common to obesity and substance use disorders in a large human population.

The hypothesis that drugs and certain nutrients found in food compete for overlapping reward mechanisms is an intuitively appealing explanation for the lack of epidemiological studies reporting comorbidity between obesity and substance use disorders. Yet this hypothesis is also difficult to test with observational studies alone. Relevant lifestyle measures have complex, time-dependent relationships with one another:[60] (1) alcohol contains 7 calories per gram with inconsistent effects on food intake,[61] (2) the nicotine found in tobacco suppresses appetite and decreases food intake,[62] (3) in public environments where alcohol is readily available cigarettes are also more likely to be present,[63] and (4) commercially available foodstuffs high in fat are more likely to be high in sugar.[64] These challenges aside, a recent cross-sectional, population-based study among 6,121 Chinese adult male twin pairs found that the effects of genetics on body mass index (BMI) were less influential in those individuals currently drinking alcohol, thus demonstrating a clear gene-alcohol interaction effect on BMI.[65] The effects of gene-smoking interactions on BMI have also been reported, with one study showing a 38% increase in the variance of BMI explained when taking gene-smoking interactions into account.[66] As these studies highlight, any approach to predict obesity or substance use disorders must address the combinatorial problem of the interactions between modifiable and non-modifiable risk factors over time. How best to accomplish this goal is a nontrivial task.

Recently, Bastarache and colleagues made headway on this front after making the astute observation that genetic association studies often examine phenotypes independently, potentially

missing groups of people with multiple phenotypes that share a single cause.[67] By aggregating phenotypes together using well-characterized diseases with Mendelian inheritance patterns, they uncovered 18 associations between rare variants and phenotypes consistent with Mendelian diseases. Implicit in this approach is that a set of phenotypes can serve as a signal for a conserved biological system, but that in any given group of people this biological system can have more than one independent *genetic* cause—i.e., biological systems contain redundancies. We extend their framework here by again using a set of phenotypes to serve as a signal for a conserved biological system, but instead we allow for this biological system to have more than one independent genetic *or environmental* cause. To explore the utility of this approach, we chose the reward system as our use case with relative nutrient intake and drug use across the population as our set of phenotypes. The dual aims of the present study were (1) to uncover novel variants associated with this biological system, and (2) improve prediction of obesity and substance addiction in humans by allowing for independent reward system-genetic and environmental causes of disease.

To accomplish this, we first defined a set of reward system-related phenotypes comprising quantitative measures of daily nutrient intake and drug use (i.e., percent energy from total fat, percent energy from total sugars, milligrams of caffeine, grams of alcohol, and number of cigarettes per day) available for a subsample of participants of the UK Biobank project. Next, we aggregated these reward system-related phenotypes together by calculating a score based on each participant's relative nutrient intake and drug use. We then conducted a GWAS to identify reward system-related loci in participants of white British ancestry. Lastly, from the results of this GWAS we constructed a reward system PRS to predict participants with obesity (BMI $\geq$ 30 kg/m^2) and substance addiction. Here, each participant's substance addiction status was defined

as a self-reported affirmative history of or current addiction to a substance (but not a behavior), including alcohol, illicit or recreational drugs, and prescription or over-the-counter medications. If a participant self-reported ≥ 10 pack-years of smoking, we also designated that the participant had a substance addiction. In addition to the reward system PRS, environmental risk factors for obesity and substance addiction that could be reasonably ascertained during a pediatric clinic visit (specifically, residing in a neighborhood with a high rate of poverty and a history of physical or sexual abuse as a child) were included in our prediction models. To evaluate any improvement in predictive ability using our reward system PRS, we compared its performance to an obesity-specific PRS and a substance addiction-specific PRS—as well as null phenotype PRS—in our obesity and substance addiction models, respectively, generated using the same training, validation, and test sets.

The findings from this work include four reward system aggregate phenotype-associated loci that were significant at the genome-wide level. Two of these loci contain genes or previously identified variants associated with one of the five constituent phenotypes comprising our reward system aggregate phenotype, while the other two loci contain variants previously identified in GWAS of obesity. Our PRS prediction models of obesity and substance addiction, however, recapitulated known challenges in translating disease biology into clinically actionable insights. Specifically, neither of our obesity or substance addiction PRS performed better than a PRS generated from a set of five quantitative phenotypes chosen at random. These results suggest that while an expert-curated aggregate-phenotype approach to predict related complex genetic diseases may prove effective in specific at-risk groups (e.g. adolescents) in the future, we did not find evidence here that this approach was effective for predicting obesity and substance addiction in middle-aged all-comers.

METHODS

The statistical analysis plan dated April 23, 2019 is provided as Appendix A.

Data source

We used data donated by people who participated in the UK Biobank project.[36] Briefly, the UK Biobank project is a publicly available, controlled-access prospective cohort study of approximately 500,000 participants. UK Biobank project participants were living in the United Kingdom (UK) and between the ages of 40 and 69 years old at recruitment, which occurred between 2006 and 2010. UK Biobank participants contributed their phenotypic and genetic data (including their electronic health records) to the project by traveling to one of 22 assessment centers located throughout the UK. The UK Biobank researchers chose the locations of these centers so as to include a diverse set of participants from both urban and rural communities of various socioeconomic backgrounds. Around 6% of people contacted by the UK Biobank researchers chose to participate, and these participants were healthier on average than the general population of the UK.[68]

After giving written consent, each of the participants completed a series of surveys, provided physical measurements (e.g., body size, imaging, etc.), and donated biological samples (i.e., blood, urine, and saliva), which are now stored in Stockport, UK. Between 2013 and 2015, the UK Biobank researchers extracted DNA from the participants' blood samples and genotyped approximately 800,000 markers per participant using two custom-designed arrays (with 95% of the genotype markers common between the two arrays). A more detailed description of the characteristics of the entire UK Biobank cohort are described elsewhere.[37]

Study participants

We used a subsample of participants from the entire UK Biobank cohort for our study. To be included in our study, each UK Biobank participant had to meet the following criteria: (1) they reported their age (i.e., month and year of birth) and sex, (2) they were genotyped by the UK Biobank researchers and assigned to the white British ancestry subset, (3) they had their height and weight measured at least once, (4) they answered the requisite questions from the smoking section of the lifestyle and environment questionnaire and the addictions section of the mental health questionnaire used to determine substance addiction status, and (5) they completed at least one 24-hour dietary recall questionnaire (i.e., a date was recorded for when the diet questionnaire was completed). The total number and proportion of UK Biobank participants with complete data under these five criteria are shown in Figure 1. The final study subsample comprised 81,420 UK Biobank participants (16%). Unless explicitly stated otherwise, all data management and statistical analyses were performed using the R software for statistical computing.[69] A summary of our UK Biobank-approved research and a list of all 287 distinct data fields we received from the UK Biobank in our project application are provided as Appendix B.

Study design

Before partitioning the participants in our study subsample, we first determined the proportion of participants who were obese (BMI $\geq$ 30 kg/m^2) and who reported either a history of or a current addiction to a substance, which is shown in Table 1. Next, we split the participants in our study subsample into three groups: 70% of the participants were randomly assigned to the training set (i.e., the "GWAS discovery sample"), 15% to the validation set (i.e., the "GWAS target sample"), and 15% to the test set (i.e., "PRS validation sample"). Participants from the

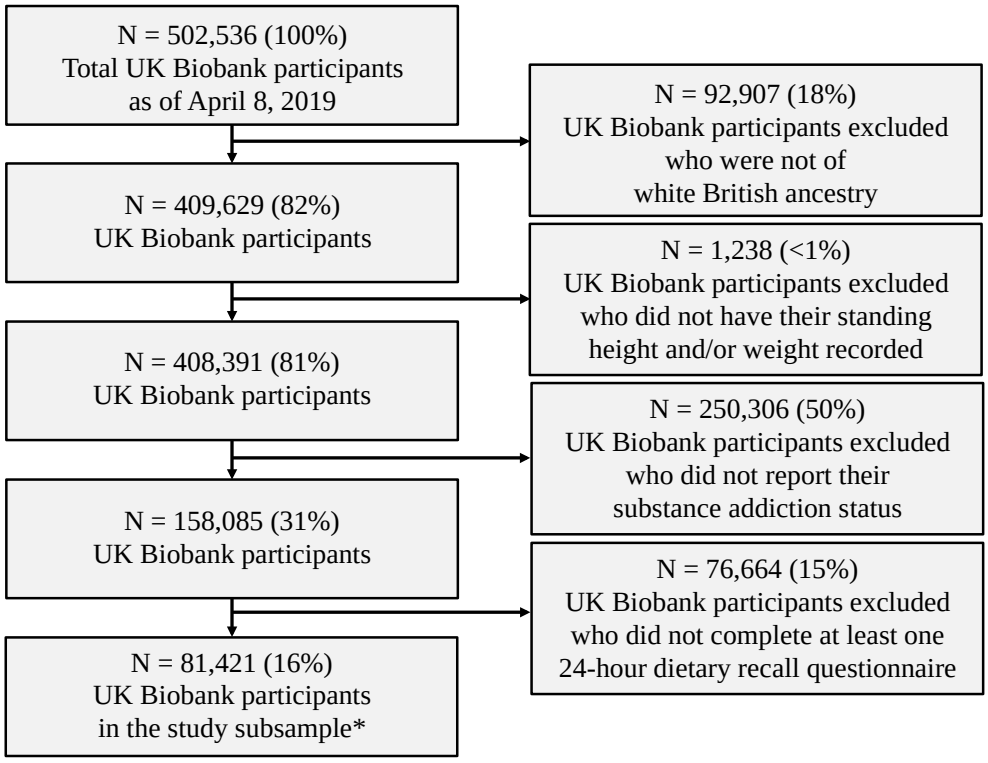


Figure 1. Derivation of the study subsample from the UK Biobank cohort based on the exclusion criteria at right. *One UK Biobank participant from the study subsample withdrew and was removed effective February 2, 2020. The final study subsample includes N = 81,420 UK Biobank participants.

study subsample were randomly assigned to these three groups such that that the relative proportion of participants with the two study outcomes (i.e., obesity and substance addiction) in each group was approximately equivalent to their relative proportions in the entire study subsample. The results of this split, along with statistics on the age and self-reported sex of the participants in the entire study subsample, the training, validation, and test sets are also shown in Table 1. The training set was then used for our genome-wide association studies, while the validation and test sets were used to fit and test our polygenic risk score prediction models, respectively (see “Genome-wide association analyses” and “Polygenic risk score analyses” sections below). An overview of the study design is shown in Figure 2.

Table 1. Comparison of study outcomes across the UK Biobank cohort, the study subsample, the training, validation and test sets.

	UK Biobank* (N = 502,536)	Study subsample* (N = 81,421)	Training set* (70%) (N = 56,994)	Validation set (15%) (N = 12,214)	Test set (15%) (N = 12,213)
Age, mean (standard deviation), years	56.5 (8.1)	56.6 (7.7)	56.6 (7.7)	56.6 (7.7)	56.6 (7.6)
Self-reported sex, number (%)					
Women	273,402 (54.4)	43,766 (53.8)	30,591 (53.7)	6,520 (53.4)	6,655 (54.5)
Men	229,134 (45.6)	37,655 (46.2)	26,403 (46.3)	5,694 (46.6)	5,558 (45.5)
Obesity, number (%)	(N = 499,520)				
BMI $\geq$ 30 kg/m^2	122,281 (24.5)	18,179 (22.3)	12,695 (22.3)	2,795 (22.9)	2,689 (22.0)
BMI < 30 kg/m^2	377,239 (75.5)	63,242 (77.7)	44,299 (77.7)	9,419 (77.1)	9,524 (78.0)
unadjusted P-value[#]			*0.816*	*0.169*	*0.443*
BMI, mean (standard deviation), kg/m^2	27.4 (4.8)	27.1 (4.7)	27.1 (4.7)	27.2 (4.7)	27.1 (4.7)
Substance addiction, number (%)	(N = 189,620)				
Substance addiction	116,617 (61.5)	36,861 (45.3)	25,636 (45.0)	5,671 (46.4)	5,554 (45.5)
No substance addiction	73,003 (38.5)	44,560 (54.7)	31,358 (55.0)	6,543 (53.6)	6,659 (54.5)
unadjusted P-value[#]			*0.283*	*0.017*	*0.673*
Joint outcomes, number (%)	(N = 188,768)				
BMI $\geq$ 30 kg/m^2 and a substance addiction	35,032 (18.6)	10,474 (12.9)	7,272 (12.8)	1,637 (13.4)	1,565 (12.8)
BMI $\geq$ 30 kg/m^2 and no substance addiction	13,202 (7.0)	7,705 (9.4)	5,423 (9.5)	1,158 (9.5)	1,124 (9.2)
BMI < 30 kg/m^2 and a substance addiction	80,892 (42.8)	26,387 (32.4)	18,364 (32.2)	4,034 (33.0)	3,989 (32.7)
BMI < 30 kg/m^2 and no substance addiction	59,642 (31.6)	36,855 (45.3)	25,935 (45.5)	5,385 (44.1)	5,535 (45.3)

*Total UK Biobank participants as of April 8, 2019. One UK Biobank participant from the study subsample (and training set) withdrew and was removed effective February 2, 2020. The final study subsample and training set include N = 81,420 and N = 56,993 UK Biobank participants, respectively. [#]No difference in study outcomes between the study subsample and the training, validation, or test set (Bonferroni correction, $P \geq 0.05 / 6$).

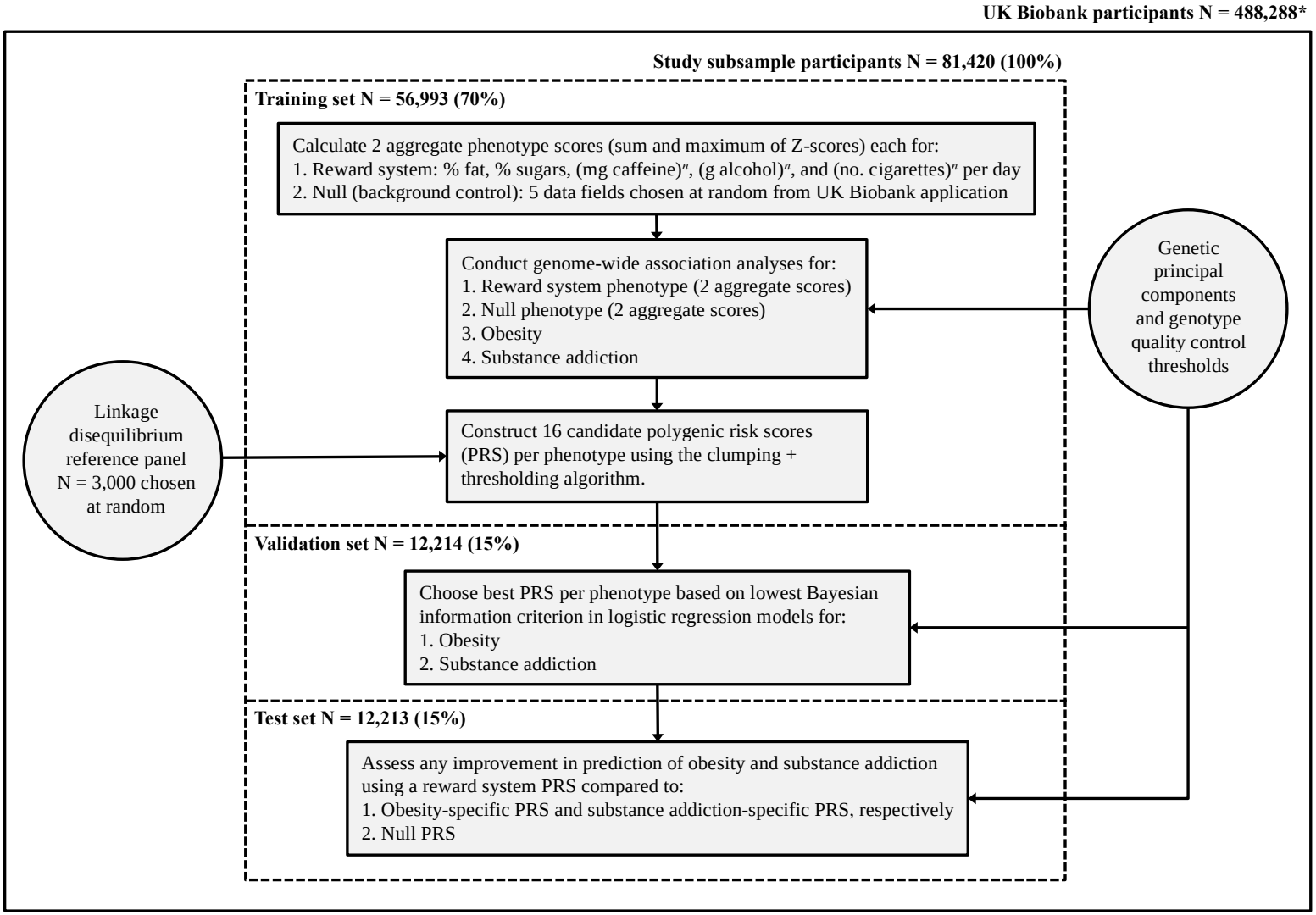


Figure 2. Study design overview, where *n* is a power transformation (see "Methods"). *Total genotyped UK Biobank participants as of April 8, 2019.

The rates of obesity and substance addiction, as well as age and self-reported sex, for all UK Biobank participants (whose data are available) are provided for comparison in Table 1. Maps showing the percentage of total participants from each geographic region in the UK Biobank cohort and our study subsample are provided as Appendix C.

Training set phenotype and environment EDA and expert review

Before constructing our reward system aggregate phenotype scores and null phenotype scores, we conducted an exploratory data analysis (EDA). First we evaluated the reliability of the study outcomes, as well as the data fields used to derive these two outcomes (i.e., BMI and pack years of smoking) across repeat assessments. For the data fields used to derive our study outcomes, we used each participant's earliest recorded value if it was measured or reported more than once. We evaluated the performance of both the smoking section of the lifestyle and environment questionnaire and the addictions section of the mental health questionnaire by calculating the proportion of questions answered by each participant who started these sections. We evaluated whether there were differences in the outcomes based on study-specific covariates (e.g., assessment center, device ID, etc.).

Second, we derived three phenotypes that we anticipated using in our reward system aggregate phenotype scores from the distinct data fields we received from the UK Biobank (see "Reward system aggregate phenotype scores" section below). To derive percent energy from total fat yesterday and percent energy from total sugars yesterday, we divided each participant's estimated total fat yesterday (g) and estimated total sugars yesterday (g) by estimated total energy yesterday (kJ) using each nutrient's energy content reference value: 37 kJ/g (9 kcal/g) fat and 17 kJ/g (4 kcal/g) sugar. This allowed us to compare relative daily nutrient intakes across

participants while accounting for differences in each participant's daily energy requirements. Here, participants' intake of total sugars includes both sugars naturally present in foods and those added during food production.[70] To derive caffeine yesterday (mg), we estimated use from coffee and tea drank by the participants using their completed 24-hour dietary recall questionnaires and the U.S. Department of Agriculture caffeine content reference values (see Appendix D for details).

Next we checked the proportion of missing data and evaluated the reliability of the data fields we anticipated using in either our reward system aggregate phenotype scores or null phenotype scores (see "Null phenotype scores" section below). We evaluated whether there were differences in these data fields based on study-specific covariates (e.g., assessment center, the number of diet questionnaires each participant completed, the hour of the day, the day of the week, and the season each participant completed the questionnaire, how long it took each participant to complete the questionnaire, etc.).

For all of the data fields we anticipated using in either our reward system aggregate phenotype scores or null phenotype scores, we examined the distributions using both visual and numerical descriptive statistical summaries and noted (but did not remove) any outliers. We examined the relationships between every pair of nutrients and drugs using both visual and numerical (i.e., Spearman's correlation coefficients) summaries, and noted any non-linear relationships.

Lastly, we checked the proportion of missing data for the data fields we anticipated using as environmental covariates in our polygenic risk score analyses (i.e., Townsend deprivation index[71] of self-reported home location at recruitment and a self-reported affirmative history of

physical or sexual abuse as a child[72]). Again, we evaluated whether there were differences in these data fields based on study-specific covariates (e.g., assessment center).

Clinical and research experts in obesity and substance use disorders reviewed the results of the EDA for measurement validity, reliability, and potential use in the reward system aggregate phenotype scores. A summary of notable EDA results is included as Appendix D.

Reward system aggregate phenotype scores

Current research suggests that chemical properties inherent to particular drugs and nutrients found in food increase the probability that a person will seek them out again; specifically, repeated intake over time induces a pathophysiological response in the brain that limits a person's capacity to avoid the particular drug or nutrient in the future.[9] While the extent to which this response in humans is similar across various self-administered chemical substances is unknown, fat and sugar are the substances found in food that consistently elicit this response.[73]

Next, we considered these nutrients found in food, together with drugs, as pharmacological agents that produce an effect on a person's reward system. We then proposed that the amount of drugs and nutrients that a person self-administers each day is an estimate of the dose required to achieve that person's individual pharmacokinetic steady state. From this assumption, it follows that the variation in this steady state dose across a human population reflects differences in the underlying biology of the human reward system.

We then calculated two aggregate phenotype scores for each participant in our training set, which represent two alternative biological models of reward system function. The two alternative ways we chose to collectively model drug use and nutrient intake in a single

participant were: (1) calculate a sum, or (2) take the maximum across all drugs and nutrients that a given participant consumes each day:

$$\text{Reward System Aggregate Phenotype Score}_{\text{S}} = \sum_{i \in \{S\}} Z_i \quad (1)$$

$$\text{Reward System Aggregate Phenotype Score}_{\text{M}} = \max_{i \in \{S\}} Z_i \quad (2)$$

Here, *Z* is a given participant's standard score (i.e., *Z*-score) as compared to the rest of the study participants in the training set, and *S* is a set containing the participant's drug use and nutrient intake per day. When a participant completed the same questionnaire multiple times, we used the participant's average daily consumption.

In our additive model above (i.e., Reward System Aggregate Phenotype Score_{S}), a higher score for a given participant represents a higher daily dose *on average across multiple nutrients and drugs* needed to achieve the pharmacokinetic steady state, while a lower score represents a lower daily dose on average across these same nutrients and drugs to achieve the steady state. Alternatively, in our single agent model above (i.e., Reward System Aggregate Phenotype Score_{M}), a higher score for a given participant represents a higher daily dose of *at least one of the nutrients or drugs* needed to achieve the pharmacokinetic steady state, while a lower score represents a lower daily dose across any of these same nutrients or drugs to achieve the steady state.

The nutrients and drugs we included in our reward system aggregate phenotype scores were: percent energy from total fat yesterday, percent energy from total sugars yesterday, caffeine yesterday (mg), alcohol yesterday (g), and number of cigarettes currently smoked daily. Cannabis and other drugs were not included in our reward system aggregate phenotype scores as quantitative data on their daily use were not available. We anticipated that average intake of each nutrient (as a proportion of total energy) yesterday would follow a normal distribution.[74]

Meanwhile, we anticipated that average drug use yesterday (and average daily drug use) would follow a Poisson distribution[75] (where $\lambda = 1$) or, alternatively, a log(x+1)-normal distribution. We also anticipated that there might be a disproportionate number of participants with very little to no average drug use per day, such that these data would be more appropriately modeled using a zero-inflated distribution.[76] Depending on the distribution of best fit (as determined by a comparison of one-sample Kolmogorov–Smirnov and Poissonness plot[77] test statistics), we anticipated performing a square root or other power transformation (estimated using the maximum likelihood-like approach of Box-Cox[78]) or a log(x+1) transformation. The test statistics used to evaluate the distribution of best fit for average nutrient intake and drug use are shown in Appendix E (Q-Q and Poissonness plots not shown). Based on these results and the previously described distributions that are theoretically appropriate to model drug use, we performed a power transformation of $^3/_5$, $^2/_5$, and $^1/_2$ to mean caffeine yesterday (mg), mean alcohol yesterday (g), and mean number of cigarettes currently smoked daily, respectively.

Similar to our assessment of the untransformed data in our EDA, we examined the visual and numerical descriptive statistical summaries of these now transformed data and noted (but did not remove) any outliers. We also examined the relationships between every pair of nutrients and drugs using both visual and numerical (i.e., Pearson correlation coefficients) summaries, and noted any non-linear relationships. A summary of notable results from the development of our reward system aggregate phenotype scores is included as Appendix E.

Null phenotype scores

We calculated two null phenotype scores to serve as background controls to our two reward system aggregate phenotype scores. To accomplish this, we first sampled five data fields

(of type continuous or integer) at random from the 287 distinct data fields we received in our UK Biobank application (see Appendix B). If any of the fields chosen at random were missing for all repeat assessments in greater than 50% of the participants, we resampled until all the chosen data fields were not missing for a majority of the participants.

Similar to the calculation of our reward system aggregate phenotype scores above, we calculated (1) a sum, and (2) the maximum across the randomly chosen data fields:

$$\text{Null Phenotype Score}_{\text{S}} = \sum_{i \in \{S\}} Z_i \quad (3)$$

$$\text{Null Phenotype Score}_{\text{M}} = \max_{i \in \{S\}} Z_i \quad (4)$$

Here, Z is a given participant's standard score (i.e., Z-score) as compared to the rest of the study participants in the training set, and S is a set containing the randomly chosen data fields. If repeat assessments were available for any of the randomly chosen data fields, we calculated each participant's average across these assessments. For the randomly chosen data fields that were not approximately normally distributed, we performed a log or power transformation to yield a distribution as close to normal as possible. The five data fields chosen at random comprising our null phenotype scores are included in Appendix E.

Genotype quality control

The UK Biobank researchers released a final dataset of 488,377 participants genotyped at 97,059,328 markers (805,426 of these markers were genotyped directly) after extensive marker and sample quality control, haplotype estimation[79] and genotype imputation;[80] a detailed description of the quality control pipeline, haplotype estimation, and genotype imputation performed by the UK Biobank researchers is described elsewhere.[37] As thresholds set by the UK Biobank researchers to designate poor quality markers and samples across the entire genotyped

cohort were not particularly stringent (so as to allow researchers to further refine the thresholds more appropriate for their own studies), we performed a secondary quality control step. We removed genotype markers that: (1) had a minor allele frequency (MAF) $\leq 0.1\%$, or (2) had an imputation quality score[81] ≤ 0.3 among the UK Biobank's final genotyped dataset. We removed participants if UK Biobank researchers: (1) determined that their samples were of poor quality (i.e., they were outliers for genotype missing rate or heterozygosity), (2) did not impute their genotypes across all chromosomes, (3) determined that the sex inferred from their genetics did not match their self-reported sex, or (4) they were a third degree relative or closer to at least one other genotyped UK Biobank participant (i.e., we used the maximal set of unrelated participants inferred by the UK Biobank researchers). In Appendix F we show the number and proportion of markers and samples we removed from the training set using these quality control criteria *based on the entire UK Biobank genotyped cohort*; the number and proportion of markers and samples that would have been removed if we had used the same quality control criteria but *based on the training set participants only* are provided for comparison. Management of the UK Biobank binary genotype files was performed using the BGEN suite of software tools.[82]

Genome-wide association analyses

After partitioning our study subsample (see "Study design" section above), we performed six GWAS using participants from the training set. The six outcome phenotypes for these analyses were: (1) Reward System Aggregate Phenotype Score_S, (2) Reward System Aggregate Phenotype Score_M, (3) Null Phenotype Score_S, (4) Null Phenotype Score_M, (5) obesity (BMI $\geq$ 30 kg/m^2), and (6) substance addiction (see "Introduction" for definition). The first four GWAS

were quantitative trait loci analyses conducted using linear regression models, while the latter two GWAS were case-control analyses conducted using logistic regression models.

The central (alternative) hypothesis of this study was that there was at least one variant genome-wide associated with *either* reward system aggregate phenotype score. The null hypothesis was that there was no relationship between any genome-wide variant and either reward system aggregate phenotype score. There was no hypothesis per se for GWAS #3-6; instead, the output of these analyses was used to evaluate any relative improvement in predictive ability of our reward system polygenic risk score models (see section "Polygenic risk score analyses" below for further details). The GWAS of obesity was considered a replication study.

To test our central hypothesis, we first built six regression base models—corresponding to the six GWAS above—simultaneously. We included age, sex, array, assessment center, and the first 9 genetic principal components (PCs) as covariates across all six base models. Additionally, we included a covariate to indicate whether a participant had their weight measured using the Tanita BC418MA body composition analyzer or a standard scale in the obesity base model. The rationale behind including 9 genetic PCs was based on an evaluation of the geographical evidence for their inclusion—i.e., we tested whether the north or east coordinates of each participant's home location at assessment could explain a significant amount of the variation in genetic PCs 1-10, after accounting for array and assessment center (two-sided Wald test based on the t-distribution, Bonferroni correction for multiple hypothesis testing where m = 20). GWAS were performed with PLINK using a linear transformation (mean of zero and variance of one) of all quantitative phenotypes and covariates. [83,84] A Firth logistic regression model was fit for any variants that failed to converge with a logistic regression model.[85]

We used the genetic PCs calculated from the entire genotyped UK Biobank cohort as covariates in our base models (as provided by the UK Biobank researchers), rather than re-calculating the genetic PCs using either the training set participants only or all UK Biobank participants in the white British ancestry subset (see Figure 2). The rationale behind this decision was based on the assumption that genetic PCs calculated from a population sample containing people of diverse ancestries (i.e., from around the world) can be used to correct for population stratification in complex disease GWAS of a homogeneous subsample of people of a single ancestry without adding noise sufficient to obscure any genome-wide significant findings. Ma and Amos provided theoretical justification for this decision.[86] In addition, we anticipated and confirmed concordance between variants significantly associated with obesity using (1) our obesity replication GWAS that employed this "population" genetic PCs approach, and (2) the results of previously published obesity GWAS that did not employ this approach.[39]

To assess whether a variant was associated with either reward system aggregate phenotype score, we used a two-sided Wald test based on the t-distribution.[84] We counted each minor allele at multi-allelic sites as independent markers. Since we performed multiple tests (one each for the total number of variants tested), we used the multiple hypothesis testing adjustment procedure proposed by John Storey with an acceptable false discovery rate set at 5%, which was partitioned evenly between the two reward system aggregate phenotype scores.[87] We confirmed that none of the reward system aggregate phenotype-associated loci could explain a significant amount of the variation in genetic PCs 1-9 (two-sided Wald test based on the t-distribution, Bonferroni correction for multiple hypothesis testing where m = 9).

Polygenic risk score analyses

We constructed PRS from the output of the six GWAS conducted on the training set above. To construct these six PRS, we applied the same genotype marker and sample quality control criteria to the validation and test sets as the training set (see above under "Genotype quality control"). Before constructing the six PRS, we used the participants from the validation set to fit two logistic regression base models (using the same covariates as our GWAS base models) to predict obesity and substance addiction. We included Townsend deprivation index of home location at recruitment and history of physical or sexual abuse as a child as environmental covariates in these logistic regression base models.

To obtain each participant's six PRS, we calculated the average of the count of their risk alleles weighted by each allele's GWAS effect size (or log odds ratio). To determine exactly which risk alleles to include in each of the six PRS, we first generated 16 candidate PRS per GWAS (for a total of 6 x 16 = 96 PRS) using the clumping and thresholding algorithm as implemented in PLINK.[83,84] For the 16 candidate PRS, we used a range of r^2-values within a 250kb sliding window (r^2 = 0.2, 0.4, 0.6, 0.8) and a range of *P*-values (*P* = 0.5, 0.05, 5×10^{-4}, 5×10^{-8}). For our linkage disequilibrium reference panel, we used a random sample of 3,000 participants from the entire genotyped UK Biobank cohort after applying the same genotype marker and sample quality control criteria as the training set (see Figure 2). The rationale behind this decision is similar to the one provided above for using the genetic PCs from the entire genotyped UK Biobank cohort—i.e., modeling linkage disequilibrium at the population level is theoretically sufficient to correct for population stratification in a (biased) sample.[86] In addition, we included variants in our PRS only if they were located on the autosomes or within the pseudo-autosomal region of the sex chromosomes. For variant sites that were multi-allelic, we

chose one minor allele at random for inclusion across all of our PRS. After generating these 16 candidate PRS, we then choose one PRS per GWAS with the lowest Bayesian information criterion (BIC) for inclusion in our obesity and substance addiction logistic regression prediction models. The logistic regression model equations for obesity were:

$$\ln\left\{\frac{Pr(\text{obesity})}{1-Pr(\text{obesity})}\right\} = \beta_0 + \beta(\text{age}) + \beta(\text{sex}) + \beta(\text{array}) + \beta_n(\text{assessment center}_{\,n=21}) + \beta_n(\text{genetic principal component}_{\,n=9}) + \beta(\text{scale type}) + \beta(\text{Townsend deprivation index}) + \beta(\text{child abuse}) + \beta(PRS_i) \quad (5)$$

where PRS_i is a polygenic risk score of obesity, the Reward System Aggregate Phenotype $Score_S$, the Reward System Aggregate Phenotype $Score_M$, the Null Phenotype $Score_S$, or the Null Phenotype $Score_M$. Likewise, the logistic regression model equations for substance addiction were:

$$\ln\left\{\frac{Pr(\text{substance addiction})}{1-Pr(\text{substance addiction})}\right\} = \beta_0 + \beta(\text{age}) + \beta(\text{sex}) + \beta(\text{array}) + \beta_n(\text{assessment center}_{\,n=21}) + \beta_n(\text{genetic principal component}_{\,n=9}) + \beta(\text{Townsend deprivation index}) + \beta(\text{child abuse}) + \beta(PRS_i) \quad (6)$$

where PRS_i is a polygenic risk score of substance addiction, the Reward System Aggregate Phenotype $Score_S$, the Reward System Aggregate Phenotype $Score_M$, the Null Phenotype $Score_S$, or the Null Phenotype $Score_M$.

Finally, we used our two reward system aggregate phenotype score PRS models to predict obesity and substance addiction among the participants in the test set. To evaluate any relative improvement in the prediction of obesity or substance addiction using a reward system

aggregate phenotype score PRS, we compared the area under their receiver operating characteristic (ROC) curves against the obesity-specific PRS model and the substance addiction-specific PRS model, respectively. We also evaluated our two reward system aggregate phenotype score PRS models against our background, or null phenotype, PRS models.

RESULTS

Genome-wide association analyses

The main characteristics of the participants in our reward system aggregate phenotype GWAS are summarized in Table 1 under "Training set," and in Appendices C-F. Briefly, 54% of the participants were women, the mean age was 57 years, 22% were obese (BMI $\geq$ 30 kg/m^2), 9% were current smokers, and 45% had a history of or current addiction to a substance (see Appendix D for a breakdown of participants by type of substance addiction). Seventy-eight percent of participants completed the entire addictions section of the UK Biobank mental health questionnaire, while 80% completed the entire smoking section of the UK Biobank lifestyle and environment questionnaire at least once. Sixty-eight percent of participants completed more than one 24-hour dietary recall questionnaire.

Among participants in our reward system aggregate phenotype GWAS (i.e., the training set), the mean estimated total energy intake yesterday was 8,877 kJ with a standard deviation (SD) of 2,584 kJ (2122 kcal, SD 618). The mean estimated percent energy intake from total fat and total sugars yesterday were 32.2 (SD 6.6) and 23.3 (SD 7.2), respectively. Meanwhile, the mean estimated caffeine use yesterday was 177.6 mg (SD 99.6), the mean estimated alcohol use yesterday was 17.7 g (SD 22.4), and the mean number of cigarettes currently smoked daily was 1.4 (SD 5.0) among all participants. Among participants who drank coffee, tea, or both (i.e., non-drinkers excluded), the mean estimated caffeine use yesterday was 189.6 mg (SD 91.1), approximately 2-8 oz. cups of filtered coffee or 4-8 oz. cups of standard (black) tea. Among participants who drank alcohol (i.e., non-drinkers excluded), the mean estimated alcohol use yesterday was 25.7 g (SD 22.8), approximately 3 standard drinks. Among participants who

smoked cigarettes (i.e., non-smokers excluded), the mean number of cigarettes currently smoked daily was 15.1 (SD 8.0).

The mean Townsend deprivation index, a relative measure of poverty by neighborhood, based on the participants' home location at recruitment was -1.7 (SD = 2.8), indicating that fewer participants on average in our reward system aggregate phenotype score GWAS lived in an impoverished neighborhood compared to the entire UK Biobank cohort (mean = -1.3, SD = 3.1). Seventy-eight of participants reported whether or not they were physically or sexually abused as a child, and (of those reporting) 24% reported in the affirmative.

We tested the association of ~20.3 million genetic variants in 39,710 participants of white British ancestry with either of our two reward system aggregate phenotype scores—a sum phenotype score (additive model) and a maximum phenotype score (single agent model). Four independent loci (MAF $\geq$ 0.01%) showed an association, one with the sum phenotype score and three with the maximum phenotype score (see Figure 3). The significance level (*q*-value), the estimated genetic effect size per copy of the minor allele (β), imputation quality (INFO score[81]), and the number of variants at the locus passing genome-wide significance (*q*-value $\leq$ 0.025 for either phenotype score) are detailed in Table 2. Another 13 additional independent rare loci (MAF < 0.01%) that showed an association with either of our two reward system aggregate phenotype scores are detailed in Appendix F.

The strongest association between either reward system aggregate phenotype score was observed between the 4p15 locus and the maximum phenotype score (i.e., the single agent model). This association was driven by rs79800723 (minor allele: C, MAF = 0.01%, β = 1.55, *q* = 0.003), a genetic variant located in long inter-genic non-protein coding RNA 1182 and 181kb

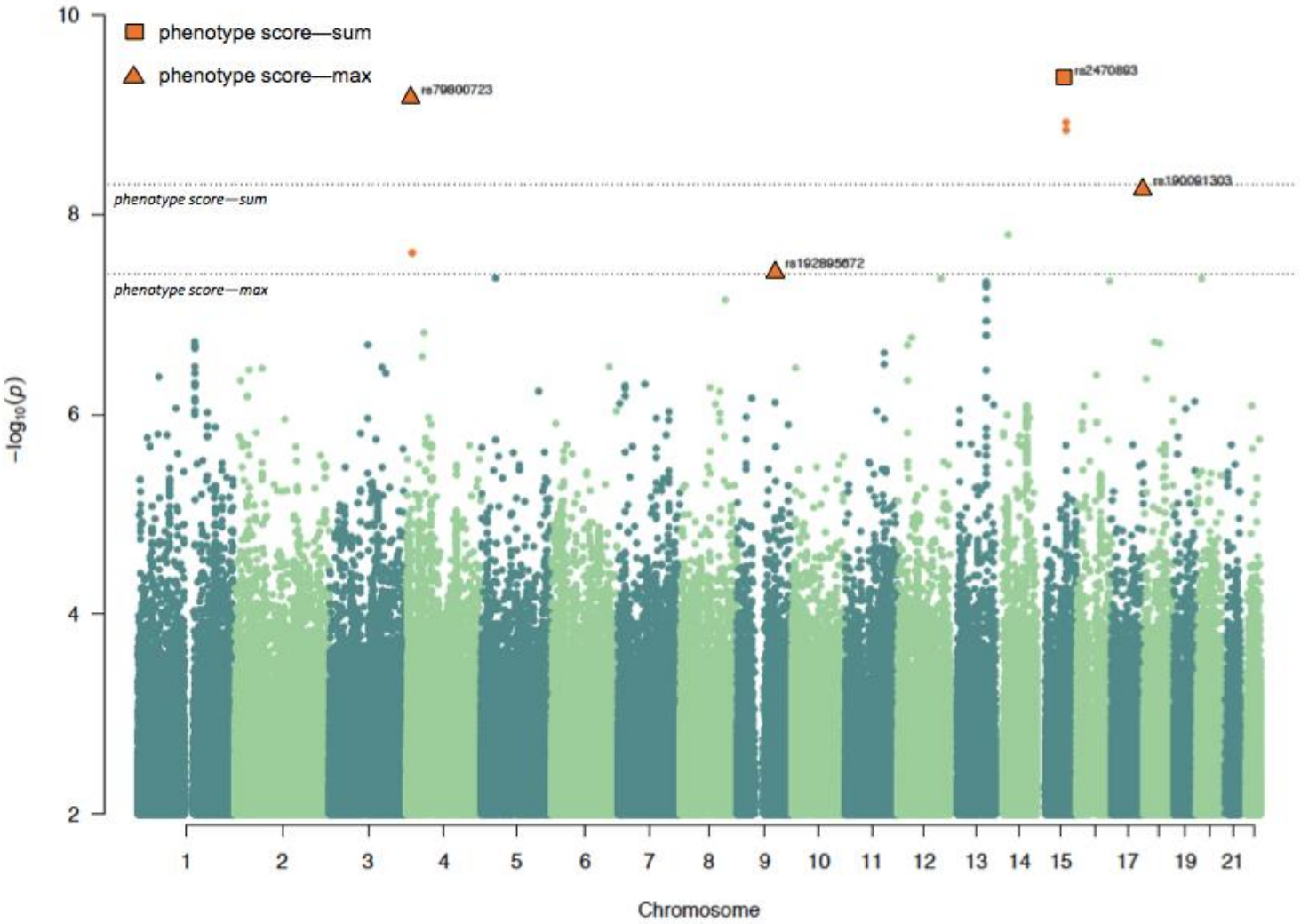


Figure 3. Manhattan plot for GWAS of reward system aggregate phenotype (MAF ≥ 0.0001). Genome-wide significant loci from analysis of sum (square) and maximum (triangle) phenotype scores combined (orange, significance line by phenotype score indicates *q*-value ≤ 0.025).

Table 2. Reward system aggregate phenotype-associated loci

Locus	Phenotype score	Chr	Locus boundaries	GWS variants at locus (no.)	Index variant	Reference/ minor alleles	MAF (UKB)	INFO score[81]	β	*q*-value	Closest gene to index variant
1	Max	4	13810020 - 13810454	2	rs79800723	T/C	0.0001 (0.002)	0.76	1.55	0.003	*BOD1L1*
2	Max	9	96709777	1	rs192895672	C/A	0.004 (0.005)	0.97	0.25	0.024	*BARX1*
3	Sum	15	74817689 - 75027880	3	rs2470893	C/T	0.331 (0.315)	1.00	0.09	0.007	*CYP1A1*
4	Max	17	79844997	1	rs190091303	G/T	0.0001 (0.001)	0.81	1.51	0.008	*ALYREF*

*Max, maximum; Chr., chromosome; GWS, genome-wide significant (*q*-value ≤ 0.025 for either phenotype score); MAF, minor allele frequency; UKB, UK Biobank. GWS variant with smallest *q*-value at locus is the index variant. Exact β and *q*-values for GWS variants with MAF < 0.0005 should be interpreted with caution. See Appendix F for rare GWS variants (MAF < 0.0001). Reference assembly GRCh37/hg19.

from the closest coding gene, *BOD1L1*. The tissue with the highest median expression of *BOD1L1* is the cerebellum.[88] Other nearby coding genes include *NKX3-2* (264kb upstream) and RAB28 (324kb upstream). Nearby genetic variants, rs73237428 (151kb), rs10939485 (271kb), and rs59341143 (280kb) have been previously identified in GWAS of weight,[89] height,[90] and hip shape as measured by dual energy x-ray absorptiometry.[91]

The second strongest association was detected between the 15q24 locus and the sum phenotype score (i.e., the additive model). This association was driven by rs2470893 (minor allele: T, MAF = 33.1%, $\beta = 0.09$, $q = 0.007$), another inter-genic single-nucleotide polymorphism (SNP) located 8kb and 22kb upstream from the genes that code for cytochrome P450 1A1 and 1A2, respectively. Representative SNP cluster plots for rs2470893 are provided in Appendix F. Both enzymes are mono-oxygenases that catalyze many reactions responsible for xenobiotic drug metabolism, in addition to the metabolism of endogenous substrates including polyunsaturated fatty acids.[92,93] CYP1A1 is primarily located in extra-hepatic tissues, while the tissue with the highest median expression of *CYP1A2* is the liver.[88] Evidence for a brain-specific *CYP1A1* splice variant that differs from the hepatic form expressed within the same individuals has also been reported.[94] Furthermore, both enzymes are inducible by polycyclic aromatic hydrocarbons, which are found in cigarette smoke. This SNP has been previously identified in GWAS of caffeine metabolism and kidney function,[59,95] and is an eQTL of *CYP1A1* expression in skeletal muscle.[88] The other two genome-wide significant variants at this locus have been previously identified in GWAS of visceral adiposity[96] (rs2472297), alcohol consumption[20] (rs2472297), and breakfast skipping[97] (rs35107470). The other genes located within this locus are *ARID3B*, *CLK3*, and *EDC3*.

The third and fourth strongest associations were detected between the maximum phenotype score and the loci 9q22 and 17q25, respectively. Only one variant at each locus reached genome-wide significance: rs192895672 at 9q22 (minor allele: A, MAF = 0.4%, β = 0.25, q = 0.024) and rs190091303 17q25 (minor allele: T, MAF = 0.01%, β = 1.51, q = 0.008). The closest gene to rs192895672 is *BARX1* (6kb downstream). The tissues with the highest median expression of *BARX1* are the stomach and the esophogus.[88] Nearby genetic variants, rs143581991 (1.5kb), rs56285369 (2.5kb), rs1933683 (3kb), and rs11789015 (6kb) have been previously identified in GWAS of bone mineral density,[98] waist-to-hip ratio[90], pyloric stenosis,[99] and digestive system disease,[100] respectively. The closest genes to rs190091303 are *ALYREF* (4.5kb downstream), *ANAPC11* (4.6kb upstream), and *NPB* (and 15kb upstream). Widely expressed in the central nervous system, the *NPB* gene codes for neuropeptide B, which modulates feeding behavior, regulates the release of corticosterone, prolactin and growth hormone while also manipulating the pain pathway.[101] There are no nearby genetic variants identified in previous GWAS at this locus.

The remaining associations between either reward system aggregate phenotype score and any genome-wide variants are provided in Appendix F. As the MAF of these variants fell below 0.01% (i.e., minor allele count < 8 among 39,710 participants), we do not elaborate on them here.

Polygenic risk score analyses

We set out to test the ability of the reward system aggregate phenotype-associated loci from the GWAS above to improve prediction of obesity and substance addiction in an independent “hold out” set comprising the last 15% of our study subsample (i.e., the test set). To

accomplish this, we first chose the best-performing sum and maximum reward system aggregate phenotype score PRS for predicting obesity among participants in the validation set using the lowest model BIC as our criterion. The model parameters for the best-performing sum reward system phenotype score PRS were $r^2 = 0.4$ and $P = 0.5$, while the model parameters for the best-performing maximum reward system phenotype score PRS were $r^2 = 0.2$ and $P = 5 \times 10^{-8}$. To assess any improvement in the prediction of obesity based on these reward system aggregate phenotype score PRS, we tested their performance against: (1) an obesity PRS, and (2) a sum and a maximum null phenotype score PRS. The model parameters of the best-performing obesity, sum and maximum null phenotype score PRS were $r^2 = 0.8$ and $P = 0.5$, $r^2 = 0.6$ and $P = 5 \times 10^{-4}$, and $r^2 = 0.2$ and $P = 0.05$, respectively. The ROC curves plotted in Figure 4 (top) show the relative ability of these five logistic regression models to predict obesity among participants in the test set. The model containing the obesity PRS as a predictor had the highest area under the ROC (AUROC) curve across the full range of possible test sensitivity and specificity thresholds (0.599, 95% CI: 0.581-0.616).[102] Neither the sum (AUROC curve: 0.574, 95% CI: 0.556-0.592) nor the maximum (AUROC curve: 0.576; 95% CI: 0.559-0.594) reward system aggregate phenotype score PRS were able to improve the prediction of obesity among participants in the test set, as both of these PRS performed similarly to the null phenotype score PRS. The AUROC curve of the model containing the sum null phenotype score PRS was 0.579 (95% CI: 0.561-0.597), while the AUROC curve of the model containing the maximum null phenotype score was 0.577 (95% CI: 0.559-0.595).

Likewise, we chose the best-performing reward system aggregate phenotype score PRS for predicting substance addiction among participants in the validation set ($r^2 = 0.8$ and $P = 0.5$ for both PRS). To assess any improvement in the prediction of substance addiction based on

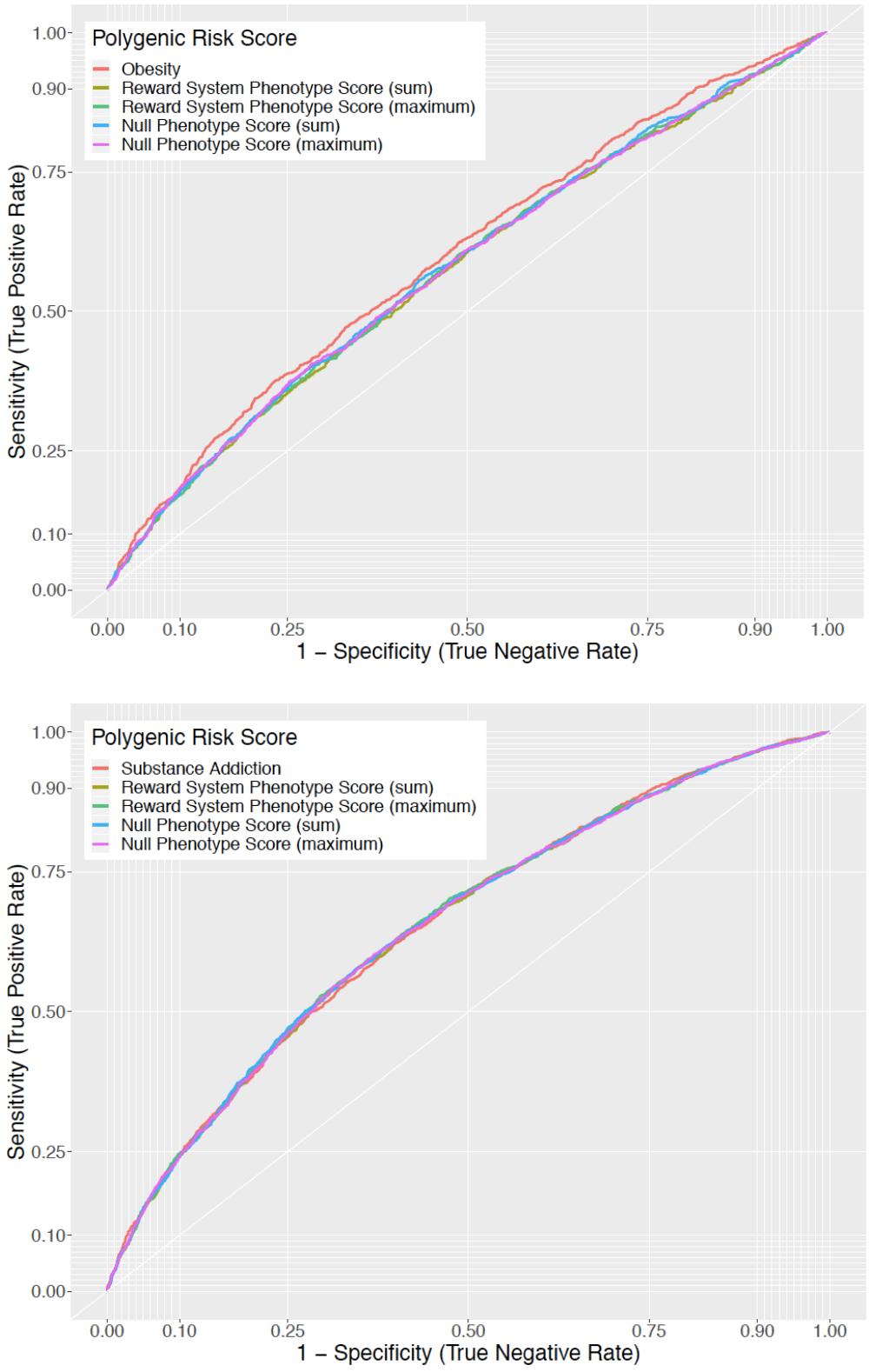


Figure 4. Receiver operating characteristic curves of polygenic risk score prediction models for obesity (top) and substance addiction (bottom).

these reward system aggregate phenotype score PRS, we tested their performance against: (1) a substance addiction PRS, and (2) a sum and a maximum null phenotype score PRS. The model parameters of the best-performing substance addiction, sum and maximum null phenotype score PRS were $r^2 = 0.8$ and $P = 0.5$, $r^2 = 0.8$ and $P = 0.5$, and $r^2 = 0.2$ and $P = 5 \times 10^{-4}$, respectively. The ROC curves plotted in Figure 4 (bottom) show the relative ability of these five logistic regression models to predict substance addiction among participants in the test set. All five PRS performed similarly across the full range of possible test sensitivity and specificity thresholds. Neither the sum (AUROC curve = 0.654; 95% CI: 0.639-0.668) nor the maximum (AUROC curve = 0.656; 95% CI: 0.641-0.670) reward system aggregate phenotype score PRS were able to improve the prediction of substance addiction among participants in the test set, as both of these PRS performed similarly to the null phenotype score PRS. The AUROC curve of the model containing the sum null phenotype score PRS was 0.655 (95% CI: 0.641-0.669), while the AUROC curve of the model containing the maximum null phenotype score was 0.654 (95% CI: 0.640-0.668).

While the higher predictive ability of the obesity PRS compared to the performance of all four phenotype score PRS suggests a lack of shared biology between nutrient intake and drug use in humans of white British ancestry, the difference in the predictive ability between all five curves is not statistically significant. Furthermore, the performance of all five PRS for predicting substance addiction were comparable, with the substance addiction PRS providing no predictive ability above a background control PRS (i.e., the null phenotype score PRS). As we were unable to generate population-level PRS models from the UK Biobank sample to predict obesity and substance addiction independently above background control PRS models, we therefore cannot

provide evidence either for or against the hypothesis that the biology underlying nutrient intake and drug use in humans is shared.

DISCUSSION

Here we described a systematic approach to test whether a set of phenotypes constitute a single biological system with more than one independent genetic or environmental cause using cross-sectional, population-based data. We incorporated information from ~20.3 million genetic variants and two well-known environmental risk factors to predict disease outcomes (i.e., obesity and substance addiction) by aggregating phenotypes comprising the reward system together into a single score. With this aggregate phenotype score, we identified four genome-wide significant loci associated with reward system function across a population of 39,710 people of white British ancestry. We then tested whether these reward system aggregate phenotype-associated loci could improve prediction of obesity and substance addiction in middle-aged adults (as compared to obesity- and substance addiction-associated loci, respectively). While the reward system aggregate phenotype-associated loci were unable to improve disease prediction, their predictive ability was approximately equivalent to that of obesity- and substance addiction-associated loci, as well as loci associated with a background control phenotype. Our results, while inconclusive with respect to the shared biology underlying nutrient intake and drug use in humans, suggest that the etiology of obesity and substance addiction may differ across the lifespan.

In earlier work, Khera and colleagues reported using 2.1 million loci from previously published GWAS of BMI to predict obesity across the lifespan from birth to middle age.[18] While they observed a 12.3 kg gradient in weight at 18 years of age between the top and bottom 10% of people by PRS, the magnitude of this gradient remained largely unchanged (Δ = 13.0 kg) when they performed the same evaluation in a separate cohort of middle-aged adults. Despite that differences in weight at the population level did not continue to diverge into adulthood, their

PRS of BMI successfully identified young adults (mean age 28.0 years) with a baseline BMI in the normal range (mean 24.2 kg/m^2) who would later go on to develop severe obesity. These findings, together with the results from the study presented here, highlight that while polygenic risk for obesity in middle age overlaps to a degree with polygenic risk for obesity in adolescents, they are not equivalent. As our study does not suggest that reward system aggregate phenotype-associated loci contribute to obesity in middle age, testing whether these loci predict obesity in adolescents is an important future direction of this work.

Likewise, previous studies of polygenic risk for substance addiction, including cigarette smoking and alcohol consumption, report differences between initiation and subsequent regular use. In a longitudinal cohort study, Belskey and colleagues tested whether a polygenic risk score generated from an earlier GWAS of cigarettes smoked per day was associated with smoking initiation and greater cigarette use over the 4-decade study period. While their polygenic risk score was associated with faster conversion from initiation to both daily and heavy smoking, it was unrelated to smoking initiation.[103] That said, a more recent analysis of drug use among 1.2 million people reported a genetic correlation between cigarettes smoked per day and age of smoking initiation that was significant after a Bonferroni correction for multiple hypothesis testing.[20] However, the magnitude of this correlation (-0.38) was smaller than the genetic correlation between age of smoking initiation and regular use (-0.71), suggesting that while the genetics that predispose people to smoking more cigarettes per day overlaps with the genetics that confer risk for smoking initiation, they are not equivalent. In a similar vein, a previous study by Sartor and colleagues reported a moderate to high positive genetic correlation between alcohol initiation and greater alcohol use (0.59), but this overlap was not complete.[104] These earlier findings, together with the results from the study presented here, warrant a future

investigation into whether our reward system aggregate phenotype-associated loci contribute to risk of heavier drug use among adolescents or young adults rather than substance addiction in adults of middle age.

There are some important limitations to the study presented here. First, the cohort used for this research was composed of individuals of white British ancestry rather than an unbiased sample of the human population. Thus, further validation is needed to confirm whether our reward system aggregate phenotype-associated loci will generalize to people of other ancestries. Despite this limitation, we provided empirical support for the "population" formulation approach proposed by Ma and Amos to account for population stratification in our polygenic risk score prediction models.[86] Such findings should reassure other researchers of the merit of this approach and facilitate ease of future PRS modeling efforts for other phenotypes of clinical importance.

A second limitation of this study is that we only tested the predictive ability of our reward system aggregate phenotype-associated loci on one easily obtainable measure of obesity (i.e., BMI $\geq$ 30 kg/m^2), and did not test the ability of these loci to predict abdominal obesity (e.g., waist-to-hip ratio, waist circumference, etc.). In light of recent reports of a possible stronger association between abdominal obesity and all-cause mortality, testing of these loci is of pressing clinical importance.[105] As our genome-wide significant reward system-associated loci contained variants previously identified in GWAS of visceral adiposity, such a study seems particularly pertinent. Likewise, we only tested the predictive ability of our reward system aggregate phenotype-associated loci on one simplified measure of substance addiction: a combination of direct, explicit questions ("Have you been addicted to…?" "Is this addiction ongoing?") along with each participant's smoking history in pack-years. This assessment does not take into account participant responses to previously validated surveys and clinical screening tools used to

identify patients at risk for substance use disorders, such as the Alcohol Use Disorders Identification Test. Given the relatively high number of participants in this study who were designated as having a substance addiction due to their smoking history and the relatively small number of participants who self-identified as having a substance addiction due to their use of other drugs (including alcohol), we must emphasize that the substance addiction outcome measure used here likely differs substantially from the diagnosis of a substance use disorder by a trained clinical professional.

Third, in formulating our models of reward system function in humans, we made the simplifying assumption that a person's self-reported, self-administered daily dose of a given drug (or nutrient) could be used to estimate that individual's pharmacokinetic steady state dose. This approach overlooks the role of alternative mechanisms for differences in drug use across humans—e.g., variation in a drug's rate of onset. Furthermore, the reward system phenotypes as well the substance addiction outcome measure (excluding height and weight) were collected via surveys rather than direct measurement. While not inherently problematic, our analysis revealed clear differences in the rates of substance addiction (and to a lesser degree the rates of obesity) among participants depending on the number of 24-hour dietary recall questionnaires they completed. One possible explanation for this finding is that the 24-hour dietary recall questionnaire did not offer participants the option to skip questions pertaining to their alcohol use (as a lack of information on alcohol use would invalidate estimates of total energy intake). While alternative estimates of alcohol use as reported by UK Biobank participants on a separate alcohol-specific questionnaire (which offered participants the option "Do not know" or "Prefer not to answer" for each question) were similar, they were not identical.

Our analysis also revealed, somewhat surprisingly, that almost all participants who did not report whether they were physically or sexually abused as a child also reported having a substance addiction (98%); these participants either specifically skipped the questions pertaining to physical or sexual abuse as a child or declined to complete the mental health questionnaire in its entirety. We highly encourage any researchers who are planning future population-level investigations into the biology common to obesity and substance addiction to allocate appropriate resources during study design to mitigate the impact of this particular source of bias and missing data.

In conclusion, we identified four loci associated with a reward system aggregate phenotype at a genome-wide significance level. Similar to previous GWAS of cigarette smoking and alcohol consumption that identified loci containing genes involved in substance-specific metabolism, the loci identified in this study contained genes broadly responsible for both polyunsaturated fatty acid and xenobiotic drug metabolism (i.e., *CYP1A1* and *CYP1A2*). Our results encourage a more thorough investigation of Phase I drug metabolic enzyme genetics and their potential role in the etiology of obesity and substance addiction. In addition, we provide an initial framework for combining genetic, environmental, and lifestyle data to improve prediction of clinical outcomes while facilitating greater understanding of human biological systems at the population level. Furthermore, our study of 81,420 people suggests that if we aim to use polygenic risk scores in the clinic for stratified prevention of obesity, substance use disorders, or their sequelae for generations to come, highly coordinated phenotypic and genetic data collection efforts and the cooperation of millions of people worldwide will be essential prerequisites.

REFERENCES

1. Khera, A.V., *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* **50**, 1219-1224 (2018).

2. Pearl, J. Does Obesity Shorten Life? Or is it the Soda? On Non-manipulable Causes. *Journal of Causal Inference* (2018).

3. Hollands, G.J., *et al.* The impact of communicating genetic risks of disease on risk-reducing health behaviour: systematic review with meta-analysis. *BMJ* **352**, i1102 (2016).

4. Rosenberg, N.A., Edge, M.D., Pritchard, J.K. & Feldman, M.W. Interpreting polygenic scores, polygenic adaptation, and human phenotypic differences. *Evol Med Public Health* **2019**, 26-34 (2019).

5. Chatterjee, N., Shi, J. & Garcia-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat Rev Genet* **17**, 392-406 (2016).

6. Torkamani, A., Wineinger, N.E. & Topol, E.J. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet* **19**, 581-590 (2018).

7. Bogdan, R., Baranger, D.A.A. & Agrawal, A. Polygenic Risk Scores in Clinical Psychology: Bridging Genomic Risk to Individual Differences. *Annu Rev Clin Psychol* **14**, 119-157 (2018).

8. González-Muniesa, P., *et al.* Obesity. *Nature Reviews Disease Primers* **3**, 17034 (2017).

9. Koob, G.F. & Volkow, N.D. Neurobiology of addiction: a neurocircuitry analysis. *Lancet Psychiatry* **3**, 760-773 (2016).

10. Albert J. Stunkard, T.T.F., Zdenek Hrubec,. A Twin Study of Human Obesity. *JAMA* **256**, 51-54 (1986).

11. Stunkard, A.J., *et al.* An adoption study of human obesity. *N Engl J Med* **314**, 193-198 (1986).

12. Kaij, L. *Alcoholism in Twins: Studies on Etiology and Sequels of Abuse of Alcohol*, (Almqvist and Wiksell International, Stockholm, 1960).

13. Goodwin, D.W., Schulsinger, F., Hermansen, L., Guze, S.B. & Winokur, G. Alcohol problems in adoptees raised apart from alcoholic biological parents. *Arch Gen Psychiatry* **28**, 238-243 (1973).

14. Madden, P.A., *et al.* The genetics of smoking persistence in men and women: a multicultural study. *Behav Genet* **29**, 423-431 (1999).

15. Merete Osler, C.H., Eva Prescott, Thorkild I.A. Sørensen. Influence of Genes and Family Environment on Adult Smoking Behavior Assessed in an Adoption Study. *Genetic Epidemiology* **21**, 193-200 (2001).

16. Thorgeirsson, T.E., *et al.* A common biological basis of obesity and nicotine addiction. *Transl Psychiatry* **3**, e308 (2013).

17. Munn-Chernoff, M.A., *et al.* Shared genetic risk between eating disorder- and substance-use-related phenotypes: Evidence from genome-wide association studies. *Addict Biol*, e12880 (2020).

18. Khera, A.V., *et al.* Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. *Cell* **177**, 587-596.e589 (2019).

19. Loos, R.J.F. & Janssens, A. Predicting Polygenic Obesity Using Genetic Information. *Cell Metab* **25**, 535-543 (2017).

20. Liu, M., *et al.* Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat Genet* **51**, 237-244 (2019).

21. Obesity and Overweight Fact Sheet. (World Health Organization, 3 March 2020).

22. Alcohol Fact Sheet. (World Health Organization, 21 September 2018).

23. Tobacco Fact Sheet. (World Health Organization, 26 July 2019).

24. Simon, G.E., *et al.* Association between obesity and psychiatric disorders in the US adult population. *Arch Gen Psychiatry* **63**, 824-830 (2006).

25. Speakman, J.R. A nonadaptive scenario explaining the genetic predisposition to obesity: the "predation release" hypothesis. *Cell Metab* **6**, 5-12 (2007).

26. Hagen, E.H., Roulette, C.J. & Sullivan, R.J. Explaining human recreational use of 'pesticides': The neurotoxin regulation model of substance use vs. the hijack model and implications for age and sex differences in drug consumption. *Front Psychiatry* **4**, 142 (2013).

27. Stephen J. Simpson, D.R. *The Nature of Nutrition: A Unifying Framework from Animal Adaptation to Human Obesity*, (Princeton University Press, 2012).

28. Merino, J., *et al.* Genome-wide meta-analysis of macronutrient intake of 91,114 European ancestry participants from the cohorts for heart and aging research in genomic epidemiology consortium. *Mol Psychiatry* (2018).

29. Clarke, T.K., *et al.* Genome-wide association study of alcohol consumption and genetic overlap with other health-related traits in UK Biobank (N=112 117). *Mol Psychiatry* **22**, 1376-1384 (2017).

30. Meddens, S.F.W., *et al.* Genomic analysis of diet composition finds novel loci and associations with health and lifestyle. *bioRxiv*, 383406 (2018).

31. BonDurant, L.D. & Potthoff, M.J. Fibroblast Growth Factor 21: A Versatile Regulator of Metabolic Homeostasis. *Annu Rev Nutr* **38**, 173-196 (2018).

32. Majuri, J., *et al.* Dopamine and Opioid Neurotransmission in Behavioral Addictions: A Comparative PET Study in Pathological Gambling and Binge Eating. *Neuropsychopharmacology* **42**, 1169-1177 (2017).

33. Volkow, N.D., Wang, G.J., Fowler, J.S. & Telang, F. Overlapping neuronal circuits in addiction and obesity: evidence of systems pathology. *Philos Trans R Soc Lond B Biol Sci* **363**, 3191-3200 (2008).

34. Uher, R. & Treasure, J. Brain lesions and eating disorders. *J Neurol Neurosurg Psychiatry* **76**, 852-857 (2005).

35. Naqvi, N.H., Rudrauf, D., Damasio, H. & Bechara, A. Damage to the insula disrupts addiction to cigarette smoking. *Science* **315**, 531-534 (2007).

36. 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).

37. Bycroft, C., *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).

38. Malin, B., Loukides, G., Benitez, K. & Clayton, E.W. Identifiability in biobanks: models, measures, and mitigation strategies. *Hum Genet* **130**, 383-392 (2011).

39. Goodarzi, M.O. Genetics of obesity: what genetic association studies have taught us about the biology of obesity and its complications. *Lancet Diabetes Endocrinol* **6**, 223-236 (2018).

40. Agrawal, A., *et al.* The genetics of addiction-a translational perspective. *Transl Psychiatry* **2**, e140 (2012).

41. Frayling, T.M., *et al.* A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**, 889-894 (2007).

42. Bierut, L.J., *et al.* Novel genes identified in a high-density genome wide association study for nicotine dependence. *Hum Mol Genet* **16**, 24-35 (2007).

43. Bierut, L.J., *et al.* A genome-wide association study of alcohol dependence. *Proc Natl Acad Sci U S A* **107**, 5082-5087 (2010).

44. Gage, S.H., Davey Smith, G., Ware, J.J., Flint, J. & Munafo, M.R. G = E: What GWAS Can Tell Us about the Environment. *PLoS Genet* **12**, e1005765 (2016).

45. Wray, N.R., Kemper, K.E., Hayes, B.J., Goddard, M.E. & Visscher, P.M. Complex Trait Prediction from Genome Data: Contrasting EBV in Livestock to PRS in Humans: Genomic Prediction. *Genetics* **211**, 1131-1141 (2019).

46. Kong, A., *et al.* The nature of nurture: Effects of parental genotypes. *Science* **359**, 424-428 (2018).

47. O'Connor, L.J. & Price, A.L. Distinguishing genetic correlation from causation across 52 diseases and complex traits. *Nat Genet* (2018).

48. Li, J., Li, X., Zhang, S. & Snyder, M. Gene-Environment Interaction in the Era of Precision Medicine. *Cell* **177**, 38-44 (2019).

49. Tyrrell, J., *et al.* Gene-obesogenic environment interactions in the UK Biobank study. *Int J Epidemiol* **46**, 559-575 (2017).

50. Lobstein, T., *et al.* Child and adolescent obesity: part of a bigger picture. *Lancet* **385**, 2510-2520 (2015).

51. Bellis, M.A., Lowey, H., Leckenby, N., Hughes, K. & Harrison, D. Adverse childhood experiences: retrospective study to determine their impact on adult health behaviours and health outcomes in a UK population. *J Public Health (Oxf)* **36**, 81-91 (2014).

52. Denny, J.C., *et al.* The "All of Us" Research Program. *N Engl J Med* **381**, 668-676 (2019).

53. Liu, B., *et al.* Development and evaluation of the Oxford WebQ, a low-cost, web-based method for assessment of previous 24 h dietary intakes in large-scale prospective studies. *Public Health Nutr* **14**, 1998-2005 (2011).

54. Rankinen, T. & Bouchard, C. Genetics of food intake and eating behavior phenotypes in humans. *Annu Rev Nutr* **26**, 413-434 (2006).

55. Tanaka, T., *et al.* Genome-wide meta-analysis of observational studies shows common genetic variants associated with macronutrient intake. *Am J Clin Nutr* **97**, 1395-1402 (2013).

56. Chu, A.Y., *et al.* Novel locus including FGF21 is associated with dietary macronutrient intake. *Hum Mol Genet* **22**, 1895-1902 (2013).

57. Thorgeirsson, T.E., *et al.* Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. *Nat Genet* **42**, 448-453 (2010).

58. Cornelis, M.C., *et al.* Genome-wide meta-analysis identifies six novel loci associated with habitual coffee consumption. *Mol Psychiatry* **20**, 647-656 (2015).

59. Cornelis, M.C., *et al.* Genome-wide association study of caffeine metabolites provides new insights to caffeine metabolism and dietary caffeine-consumption behavior. *Hum Mol Genet* **25**, 5472-5482 (2016).

60. McCarty, C.A., *et al.* Longitudinal associations among depression, obesity and alcohol use disorders in young adulthood. *Gen Hosp Psychiatry* **31**, 442-450 (2009).

61. Traversy, G. & Chaput, J.P. Alcohol Consumption and Obesity: An Update. *Curr Obes Rep* **4**, 122-130 (2015).

62. Courtemanche, C., Tchernis, R. & Ukert, B. The effect of smoking on obesity: Evidence from a randomized trial. *J Health Econ* **57**, 31-44 (2018).

63. Verplaetse, T.L. & McKee, S.A. An overview of alcohol and tobacco/nicotine interactions in the human laboratory. *Am J Drug Alcohol Abuse* **43**, 186-196 (2017).

64. Darmon, N., Ferguson, E.L. & Briend, A. A cost constraint alone has adverse effects on food selection and nutrient density: an analysis of human diets by linear programming. *J Nutr* **132**, 3764-3771 (2002).

65. Liao, C., *et al.* The association of cigarette smoking and alcohol drinking with body mass index: a cross-sectional, population-based study among Chinese adult male twins. *BMC Public Health* **16**, 311 (2016).

66. Justice, A.E., *et al.* Genome-wide meta-analysis of 241,258 adults accounting for smoking behaviour identifies novel loci for obesity traits. *Nat Commun* **8**, 14977 (2017).

67. Bastarache, L., *et al.* Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. *Science* **359**, 1233-1239 (2018).

68. Fry, A., *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol* **186**, 1026-1034 (2017).

69. Team, R.C. R: A language and environment for statistical computing. (R Foundation for Statistical Computing, Vienna, Austria, 2018).

70. Erickson, J. & Slavin, J. Total, added, and free sugars: are restrictive guidelines science-based or achievable? *Nutrients* **7**, 2866-2878 (2015).

71. P. Townsend, P.P., A. Beattie. *Health and Deprivation: Inequality and the North*, (Croom Helm, Bristol, 1988).

72. Thombs, B.D., Bernstein, D.P., Ziegelstein, R.C., Bennett, W. & Walker, E.A. A brief two-item screener for detecting a history of physical or sexual abuse in childhood. *Gen Hosp Psychiatry* **29**, 8-13 (2007).

73. Hall, K.D., *et al.* Ultra-Processed Diets Cause Excess Calorie Intake and Weight Gain: An Inpatient Randomized Controlled Trial of Ad Libitum Food Intake. *Cell Metab* **30**, 67-77.e63 (2019).

74. Anderson, J.J., *et al.* Adiposity among 132 479 UK Biobank participants; contribution of sugar intake vs other macronutrients. *Int J Epidemiol* **46**, 492-501 (2017).

75. Gorelick, D.A. & McPherson, S. Improving the analysis and modeling of substance use. *Am J Drug Alcohol Abuse* **41**, 475-478 (2015).

76. Wagner, B., Riggs, P. & Mikulich-Gilbertson, S. The importance of distribution-choice in modeling substance use data: a comparison of negative binomial, beta binomial, and zero-inflated distributions. *Am J Drug Alcohol Abuse* **41**, 489-497 (2015).

77. Hoaglin, D.C. A poissonness plot. *The American Statistician* **34**, 146–149 (1980).

78. Cox, G.E.P.B.a.D.R. An Analysis of Transformations. *Journal of the Royal Statistical Society* **26**, 211-252 (1964).

79. Delaneau, O., Marchini, J. & Zagury, J.F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179-181 (2011).

80. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).

81. Lin, P., *et al.* A new statistic to evaluate imputation reliability. *PLoS One* **5**, e9697 (2010).

82. Band, G. & Marchini, J. BGEN: a binary file format for imputed genotype and haplotype data. *bioRxiv*, 308296 (2018).

83. Purcell, S., *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575 (2007).

84. Chang, C.C., *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

85. Firth, D. Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27-38 (1993).

86. Ma, J. & Amos, C.I. Theoretical formulation of principal components analysis to detect and correct for population stratification. *PLoS One* **5**(2010).

87. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**, 9440-9445 (2003).

88. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-585 (2013).

89. Tachmazidou, I., *et al.* Whole-Genome Sequencing Coupled to Imputation Discovers Genetic Signals for Anthropometric Traits. *Am J Hum Genet* **100**, 865-884 (2017).

90. Kichaev, G., *et al.* Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *Am J Hum Genet* **104**, 65-75 (2019).

91. Baird, D.A., *et al.* Identification of Novel Loci Associated With Hip Shape: A Meta-Analysis of Genomewide Association Studies. *J Bone Miner Res* **34**, 241-251 (2019).

92. Nebert, D.W. & Dalton, T.P. The role of cytochrome P450 enzymes in endogenous signalling pathways and environmental carcinogenesis. *Nat Rev Cancer* **6**, 947-960 (2006).

93. Konkel, A. & Schunck, W.H. Role of cytochrome P450 enzymes in the bioactivation of polyunsaturated fatty acids. *Biochim Biophys Acta* **1814**, 210-222 (2011).

94. Chinta, S.J., Kommaddi, R.P., Turman, C.M., Strobel, H.W. & Ravindranath, V. Constitutive expression and localization of cytochrome P-450 1A1 in rat and human brain: presence of a splice variant form in human brain. *J Neurochem* **93**, 724-736 (2005).

95. Wuttke, M., *et al.* A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat Genet* **51**, 957-972 (2019).

96. Karlsson, T., *et al.* Contribution of genetics to visceral adiposity and its relation to cardiovascular and metabolic disease. *Nat Med* **25**, 1390-1395 (2019).

97. Dashti, H.S., *et al.* Genome-wide association study of breakfast skipping links clock regulation with food timing. *Am J Clin Nutr* **110**, 473-484 (2019).

98. Kemp, J.P., *et al.* Identification of 153 new loci associated with heel bone mineral density and functional involvement of GPC6 in osteoporosis. *Nat Genet* **49**, 1468-1475 (2017).

99. Fadista, J., *et al.* Genome-wide meta-analysis identifies BARX1 and EML4-MTA3 as new loci associated with infantile hypertrophic pyloric stenosis. *Hum Mol Genet* **28**, 332-340 (2019).

100. Levine, D.M., *et al.* A genome-wide association study identifies new susceptibility loci for esophageal adenocarcinoma and Barrett's esophagus. *Nat Genet* **45**, 1487-1493 (2013).

101. Singh, G. & Davenport, A.P. Neuropeptide B and W: neurotransmitters in an emerging G-protein-coupled receptor system. *Br J Pharmacol* **148**, 1033-1041 (2006).

102. DeLong, E.R., DeLong, D.M. & Clarke-Pearson, D.L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837-845 (1988).

103. Belsky, D.W., *et al.* Polygenic risk and the developmental progression to heavy, persistent smoking and nicotine dependence: evidence from a 4-decade longitudinal study. *JAMA Psychiatry* **70**, 534-542 (2013).

104. Sartor, C.E., *et al.* Timing of first alcohol use and alcohol dependence: evidence of common genetic influences. *Addiction* **104**, 1512-1518 (2009).

105. Hotchkiss, J.W. & Leyland, A.H. The relationship between body size and mortality in the linked Scottish Health Surveys: cross-sectional surveys with follow-up. *Int J Obes (Lond)* **35**, 838-851 (2011).

106. Morris, J.A., Randall, J.C., Maller, J.B. & Barrett, J.C. Evoker: a visualization tool for genotype intensity data. *Bioinformatics* **26**, 1786-1787 (2010).

APPENDICES

Appendix A. Statistical analysis plan

Study participants

We will use data from a subsample of people who participated in the UK Biobank project. Briefly, the UK Biobank is a publicly available, controlled access prospective cohort study of approximately 500,000 participants. Participants were living in the United Kingdom and between the ages of 40 and 69 at recruitment, which occurred between 2006 and 2010. UK Biobank participants contributed their phenotypic and genetic data (including their past and future health information via electronic medical records) to the project by traveling to one of 22 assessment centers distributed throughout the UK. The UK Biobank researchers chose the locations of these centers to target a diverse set of participants from both urban and rural communities of various socioeconomic backgrounds. Around 6% of people contacted by UK Biobank researchers chose to participate, and these participants were healthier on average than the UK population.

After giving written consent, each of the participants completed a series of touchscreen questionnaires, provided physical measurements (e.g., body size, imaging, etc.), and donated biological samples (i.e., blood, urine, and saliva), which are now stored in Stockport, UK. Between 2013 and 2015, UK Biobank researchers extracted DNA from the participants' blood samples and genotyped approximately 800,000 single nucleotide variants per participant using two custom-designed arrays (95% of the genotype markers were common between the two arrays). A detailed description of the characteristics of the entire UK Biobank cohort and study design are described elsewhere.

UK Biobank participants were eligible for inclusion in our study if they (1) completed at least one optional 24-hour dietary recall questionnaire from the previous day, (2) completed the optional mental health self-assessment questionnaire, (3) had their height and weight measured, and (4) were genotyped by the UK Biobank researchers.

Reward system aggregate phenotype scores

We will develop two aggregate phenotype scores as metrics of reward system function. To accomplish this, we will first consider addictive nutrients and drugs as pharmacological agents that produce an effect on a person's reward system. We will then assume that the amount of addictive nutrients and drugs a person self-administers each day is an estimate of the dose required to achieve that person's individual pharmacokinetic steady state. Variation in this steady state dose across a human population will reflect differences in the underlying biological makeup of the human reward system.

Two ways to measure the amount of addictive substances consumed by a single person are: (1) calculate a sum, or (2) take the maximum across all addictive nutrients and drugs that person consumes per day. We will use both of these approaches, and calculate two reward system aggregate phenotype scores for each participant:

$$Reward\ System\ Aggregate\ Phenotype\ Score_S = \sum_{i \in \{S\}} Z_i \quad (1)$$

$$Reward\ System\ Aggregate\ Phenotype\ Score_M = \max_{i \in \{S\}} Z_i \quad (2)$$

Here, Z is the participant's standard score (i.e., Z-score) as compared to the remaining UK Biobank participants, and S is a set containing the participant's current daily addictive nutrient intake and drug use. If a participant completed repeated administrations of the same questionnaire across multiple days, we will use the participant's average daily consumption (so long as the participants were equally likely to complete the questionnaire across the different days of the week).

We will include all nutrients and drugs in our reward system aggregate phenotype scores if (1) previous evidence suggests that consumption of the nutrient or drug in at least some people increases the probability that s/he will seek it out again (i.e., the nutrient or drug is "addictive"), and (2) participants' daily consumption was measured on a quantitative scale (i.e., we will not include nutrients or drugs in our study if the UK Biobank researchers only ascertained whether or not a participant consumed *any* of the nutrient or drug.)

Before we establish a common scale between the addictive nutrients and drugs, we will first assess phenotype data quality. We will report the percentage of missing data and any evidence of duplicate data. We will report the consistency of repeated measures of the same self-reported addictive nutrient intake or drug use for a given participant.

Next we will perform an exploratory data analysis. We will examine and report the distributions of each substance using both visual (e.g., histogram, box plot, Q-Q plot) and numerical (e.g., mean, standard deviation, median, range, quartiles, mode, etc.) descriptive statistical summaries and report any outliers. Before normalizing or transforming the data in any way, we will examine and report the relationships between each pair of substances using both visual (e.g., scatterplot, conditional histogram) and numerical (correlation coefficients) summaries. We will comment whether any relationships may not be linear.

We anticipate that current daily intake of addictive nutrients will be normally distributed. For each addictive nutrient, we plan to normalize each participant's daily intake by the participant's total daily energy intake (which we also expect to be normally distributed). This will allow us to compare relative daily addictive nutrient intakes across participants while accounting for each participant's individual energy requirements. We will examine and report the visual and numerical descriptive statistical summaries as above but after this normalization.

We anticipate that current daily addictive drug use will be log-normally distributed. We plan to log transform current daily use of each addictive drug so that the distributions are approximately normal. If an alternative distribution (e.g., Poisson for unrestricted count data, binomial for count data with a restricted maximum) better models current daily addictive drug use, we will consider using the corresponding appropriate transformation (i.e., square root transformation, arcsine transformation) instead. We anticipate that there may be a disproportionate number of participants with very little or no current daily addictive drug use. In this case, we will also

consider modeling current daily addictive drug use with a zero-inflated or piecewise distribution and using a power (Box-Cox) transformation.

If the distributions of current daily addictive drug use are still not approximately normally distributed after applying any one of the above transformations, we will quantile (i.e., rank) normalize across all current daily addictive drug use and nutrient intakes. We will examine and report the visual and numerical descriptive statistical summaries as above but after we have both normalized and transformed (or quantile normalized) the data. We will again examine and report the relationships between each substance using both visual and numerical (correlation coefficients) summaries. And again, we will comment whether any relationships may not be linear.

In the unlikely case that one or more measures of drug use follow a bi-modal or tri-modal distribution (suggesting that the phenotype is not actually a quantitative trait), we will conduct a case-control genome-wide association analysis for addiction to any nutrient (defined as intake in the top 5% of participants) or drug (see below under “Study design" for definition of “substance addiction”).

Study design

We will partition the participants in our final study subsample into three groups: 70% of the participants will be randomly assigned to the training set, 15% to the validation set, and 15% to the testing set. We will ensure that each of the three groups contains approximately the same proportion of participants who are obese (BMI $\geq$ 30) and who self-reported a current or past history of substance addiction (to alcohol, prescription or recreational drugs, or $\geq$ 10 pack-years of smoking). We will state participants’ self-reported age, gender, ethnicity, and their geographic distribution by these three groups, as well as altogether for comparison to the entire UK Biobank cohort. The training set will be used for our genome-wide association analyses, while the validation and testing sets will be used to fit and test our PRS prediction models, respectively.

Genome-wide association analyses

Before conducting any genome-wide association analyses, we will first assess genetic data quality. The UK Biobank researchers performed initial quality assurance and quality control (QA/QC), phasing, and imputation on the initial dataset of 489,212 UK Biobank participants genotyped at 812,428 markers.

Briefly, UK Biobank researchers identified poor quality genotype markers (0.97%) using samples from participants with inferred European ancestry (n = 463,844), the largest ancestral group within the cohort (94%). They tested for consistency across array (Affymetrix UK BiLEVE Axiom Array or Applied Biosystems UK Biobank Axiom Array), batch (4,700 samples were genotyped from array intensity data and together comprised one “batch,” with a total of 106 batches to genotype all samples), plate (each 96-well plate contained samples from 94 UK Biobank participants plus 2 individuals in the CEU group of the 1000 Genomes project, which served as controls), and sex. UK Biobank researchers also tested for departure from Hardy-Weinberg equilibrium (HWE) within each batch and discordance across control replicates.

For their sample QA/QC, UK Biobank researchers identified poor quality samples (0.2%) using a subset of high-quality autosomal genotype markers (n = 605,876). Poor quality samples had either an unusually high proportion of missing autosomal genotype markers or extreme heterozygosity, which can indicate DNA contamination or mixed samples (n = 968). Affymetrix inferred each participant's sex from the relative intensity of their sex chromosome genotype markers on their sex chromosomes. UK Biobank researchers then identified participants whose self-reported gender did not match the sex inferred by Affymetrix. UK Biobank researchers also identified participants whose sex chromosome karyotype is likely neither XX nor XY (n = 652). They identified duplicate samples that were not from identical twins, samples that were likely mishandled, and samples from participants who asked to be withdrawn from the project were removed (n = 835).

After UK Biobank researchers completed their genotype marker and sample QA/QC, they released a final dataset of 488,377 UK Biobank participants genotyped at 805,426 markers. They also confirmed that the allele frequencies among participants with inferred European ancestry were similar to another independent cohort of people with European ancestry (Exome Aggregation Consortium data, n = 33,370).

UK Biobank researchers performed phasing on the autosomes using SHAPEIT3 with the 1000 Genomes phase 3 data as a reference panel. They performed imputation using IMPUTE4 with the Haplotype Reference Consortium (HRC) data used as the main reference panel, followed by a second round of imputation using a merged UK10K and 1000 Genomes phase 3 reference panel, bringing the total number of testable variants to around 96 million.

As thresholds set by UK Biobank researchers to designate poor quality markers and samples across the entire genotyped cohort were not particularly stringent (so as to allow researchers to further refine the thresholds more appropriate for their own studies), we will perform secondary QA/QC. We will remove genotype markers that: (1) failed any of the UK Biobank's QA/QC tests in more than one batch, (2) are missing in >1% of the UK Biobank's final dataset, (3) have a minor allele frequency (MAF) < 0.1% in the UK Biobank's final dataset, or (4) have low imputation quality (i.e., information score < 0.3) as reported by UK Biobank researchers. We will include samples from participants only if: (1) they are not a third degree relative or closer to any other UK Biobank participants (i.e., we will use one of the maximal sets of unrelated participants inferred by UK Biobank researchers), (2) they are part of the white British ancestry subset as defined by UK Biobank researchers (n = 409,728), (3) their self-reported gender matches their sex inferred by Affymetrix, and (4) UK Biobank researchers imputed their genotypes.

After partitioning our final analysis subsample as described under "Study design" above, we will use the training set to perform four genome-wide association analyses. The outcomes for these analyses will be: (1) the reward system aggregate phenotype score—sum, (2) the reward system aggregate phenotype score—maximum, (3) obesity (BMI ≥ 30), and (4) substance addiction (see above under "Study design" for definition of "substance addiction"). The first two genome-wide association studies will be quantitative trait loci (QTL) analyses, while the second two genome-wide association studies will be case-control analyses.

The central (alternative) hypothesis of this project is that there is at least one variant genome-wide associated with *either* reward system aggregate phenotype score. The null hypothesis is that there is no relationship between any genome-wide variants and either reward system aggregate phenotype score. (There is no hypothesis per se for the two case-control genome-wide association analyses, as we will use the output of these analyses for polygenic risk score evaluation only.)

To test our central hypothesis, we will first build two linear regression models, one to predict each reward system aggregate phenotype score. (For the two case-control analyses, will build logistic rather than linear regression models.) We will include age, gender, and up to the first 10 principal components as predictors. (We will calculate principal components using the directly genotyped, not the imputed, markers and provide a geographic interpretation of the axes of variation using the participants' place of birth.) We will also include array as a predictor, as the sampling of participants for the UK Biobank Lung Exome Variant Evaluation (UK BiLEVE) study in addition to the array itself differed from the rest of the UK Biobank cohort. (Specifically, UK BiLEVE researchers recruited participants with either very high or low lung function who were non-smokers and heavy smokers, respectively.) In addition to array, we will consider including other study-specific covariates as predictors (e.g., assessment center, batch, plate, etc.).

To assess whether any genome-wide variant is associated with either reward system aggregate phenotype score, we plan to use a two-sided test (e.g., Wald test based on the t-distribution). Since we will perform multiple tests (2 outcomes × total number of genotype markers) under one central hypothesis, we plan to use the multiple testing procedure proposed by Storey, Taylor, Siegmund and Tibshirani with an acceptable false discovery rate set at 5%.

For any genotype markers called significant, we will check that the error has constant variance and is normally distributed. We will also visually inspect the cluster plots for the significant genotype markers to confirm that there are well-defined clusters for 0, 1, and 2 copies of the allele. We will report all genotype markers called significant (even if they are located close to one another on the chromosome). However, we will propose the boundaries of the ancestral chromosome segment from which the significant genotype marker(s) most likely arose as part of a post hoc analysis. We will report all genes located within these boundaries. We will also report if these boundaries contain variants significantly associated with related phenotypes from previous genome-wide association studies conducted using the UK Biobank cohort. We will use publicly available variant annotation and prioritization tools as well as manual literature review to highlight the functional relevance of any significant genotype markers.

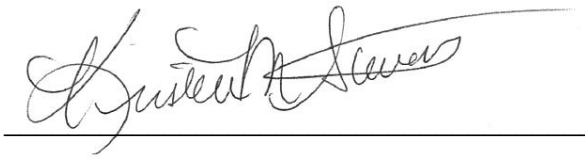
Polygenic risk scores analyses

We will construct four polygenic risk scores from the output of our four genome-wide association analyses conducted on the training set above—one for each of the reward system aggregate phenotype scores, obesity, and substance addiction. For each participant, we will calculate her/his polygenic risk scores as the sum of the count of risk alleles weighted by each allele's effect size or odds ratio.

To determine which risk alleles to include in each of our polygenic risk scores, we plan to use two approaches. We will first use the pruning and thresholding algorithm to generate 16 candidate PRSs over a range r^2-values using a 100kb sliding window (r^2 = 0.2, 0.4, 0.6, 0.8) and a range of *P*-values (*P* = 0.5, 0.05, 0.0005, 5% FDR). Second, we will use the LDPred algorithm to generate an additional 6 candidate PRSs over a range of ρ-values (ρ = 0.5, 0.05, 0.005, 0.0005), where ρ is the fraction of variants with non-zero effect sizes or odds ratios.

For each of the 20 candidate PRSs, we will choose the PRS with the best predictive ability (i.e., highest area under the receiver operator characteristic curve) across two logistic regression models, one with obesity as the outcome and one with substance addiction as the outcome. Before the predictive ability of the candidate PRSs are calculated, however, we will fit the two logistic regression base models using age, age squared, inferred sex, up to the first 10 principal components, the poverty index of the participant's postal code (i.e., Townsend Deprivation Index), and the frequency of childhood traumatic events experienced by the participant (i.e., Childhood Trauma Screener) as possible covariates. We will select the logistic regression base models with the lowest BIC.

To evaluate our two reward system polygenic risk scores, we will test their ability to predict obesity and substance addiction in participants in the testing set, again by calculating the AUROC curve. We will then compare these AUROC curves to those generated from logistic regression models using the obesity and substance addiction PRSs as predictors.



Completed by Kristen M. Stevens on April 23, 2019

Appendix B. UK Biobank approved research summary and data dictionary

biobank uk

Application number/Title: 48083 - A Novel Reward System Polygenic Risk Score for the Prediction of Obesity and Substance Use Disorders

Applicant PI: Dr Shannon McWeeney

Applicant institution: Oregon Health & Science University, Portland, Oregon, USA

Keywords provided by the Applicant PI to describe the research project: addiction, obesity, palatable food, polygenic risk score, reward system, substance use disorder

Application Lay Summary:

We currently do not know if there is a shared genetic risk for adult-onset obesity and substance use disorders. The aims of this research project are: (1) to uncover any genetic risk factors associated with reward system function, including palatable food intake and substance use, and (2) to use these genetic risk factors to improve the prediction of obesity and substance use disorders in the clinic. By investigating a common genetic signal for the reward system, this research project may improve prediction of other adult-onset complex diseases. The duration of this project will last approximately 18 months.

UK Biobank Data Dictionary for Application 48083
Date Extracted: 2019-04-08T08:53:52

*Unique Data Identifier instances, counts, and data coding removed for brevity

Distinct Data Field	Unique Data Identifier*	Type	Description
1	eid	Sequence	Encoded anonymised participant ID
2	21	Categorical (single)	Weight method
3	31	Categorical (single)	Sex
4	34	Integer	Year of birth
5	39	Text	Height measure device ID
6	40	Text	Manual scales device ID
7	41	Text	Seating box device ID
8	43	Text	Impedance device ID
9	44	Text	Tape measure device ID
10	48	Continuous	Waist circumference
11	49	Continuous	Hip circumference
12	50	Continuous	Standing height
13	51	Continuous	Seated height
14	52	Categorical (single)	Month of birth
15	53	Date	Date of attending assessment centre
16	54	Categorical (single)	UK Biobank assessment centre
17	189	Continuous	Townsend deprivation index at recruitment
18	190	Categorical (single)	Reason lost to follow-up
19	191	Date	Date lost to follow-up
20	1239	Categorical (single)	Current tobacco smoking
21	1249	Categorical (single)	Past tobacco smoking
22	1259	Categorical (single)	Smoking/smokers in household
23	1269	Integer	Exposure to tobacco smoke at home
24	1279	Integer	Exposure to tobacco smoke outside home
25	1488	Integer	Tea intake
26	1498	Integer	Coffee intake
27	1508	Categorical (single)	Coffee type
28	1538	Categorical (single)	Major dietary changes in the last 5 years
29	1548	Categorical (single)	Variation in diet
30	1558	Categorical (single)	Alcohol intake frequency.
31	1568	Integer	Average weekly red wine intake
32	1578	Integer	Average weekly champagne plus white wine intake
33	1588	Integer	Average weekly beer plus cider intake
34	1598	Integer	Average weekly spirits intake
35	1608	Integer	Average weekly fortified wine intake
36	1618	Categorical (single)	Alcohol usually taken with meals
37	1628	Categorical (single)	Alcohol intake versus 10 years previously
38	1647	Categorical (single)	Country of birth (UK/elsewhere)
39	1677	Categorical (single)	Breastfed as a baby
40	1687	Categorical (single)	Comparative body size at age 10
41	1697	Categorical (single)	Comparative height size at age 10
42	1787	Categorical (single)	Maternal smoking around birth
43	2644	Categorical (single)	Light smokers, at least 100 smokes in lifetime
44	2664	Categorical (single)	Reason for reducing amount of alcohol drunk
45	2867	Integer	Age started smoking in former smokers
46	2877	Categorical (single)	Type of tobacco previously smoked
47	2887	Integer	Number of cigarettes previously smoked daily
48	2897	Integer	Age stopped smoking
49	2907	Categorical (single)	Ever stopped smoking for 6+ months
50	2926	Integer	Number of unsuccessful stop-smoking attempts
51	2936	Categorical (single)	Likelihood of resuming smoking
52	3077	Categorical (single)	Seating box height
53	3160	Continuous	Weight, manual entry
54	3436	Integer	Age started smoking in current smokers
55	3446	Categorical (single)	Type of tobacco currently smoked
56	3456	Integer	Number of cigarettes currently smoked daily (current cigarette smokers)
57	3466	Categorical (single)	Time from waking to first cigarette
58	3476	Categorical (single)	Difficulty not smoking for 1 day
59	3486	Categorical (single)	Ever tried to stop smoking
60	3496	Categorical (single)	Wants to stop smoking
61	3506	Categorical (single)	Smoking compared to 10 years previous

Distinct Data Field	Unique Data Identifier*	Type	Description
62	3659	Integer	Year immigrated to UK (United Kingdom)
63	3731	Categorical (single)	Former alcohol drinker
64	3859	Categorical (single)	Reason former drinker stopped drinking alcohol
65	4407	Integer	Average monthly red wine intake
66	4418	Integer	Average monthly champagne plus white wine intake
67	4429	Integer	Average monthly beer plus cider intake
68	4440	Integer	Average monthly spirits intake
69	4451	Integer	Average monthly fortified wine intake
70	4462	Integer	Average monthly intake of other alcoholic drinks
71	5364	Integer	Average weekly intake of other alcoholic drinks
72	5959	Categorical (single)	Previously smoked cigarettes on most/all days
73	6157	Categorical (multiple)	Why stopped smoking
74	6158	Categorical (multiple)	Why reduced smoking
75	6183	Integer	Number of cigarettes previously smoked daily (current cigar/pipe smokers)
76	6194	Integer	Age stopped smoking cigarettes (current cigar/pipe or previous cigarette smoker)
77	6218	Integer	Impedance of whole body, manual entry
78	10115	Categorical (single)	Why stopped smoking (pilot)
79	10818	Categorical (single)	Reason for reducing amount of alcohol drunk (pilot)
80	10827	Categorical (single)	Ever stopped smoking for 6+ months (pilot)
81	10853	Categorical (single)	Reason former drinker stopped drinking alcohol (pilot)
82	10895	Categorical (single)	Light smokers, at least 100 smokes in lifetime (pilot)
83	10912	Categorical (single)	Variation in diet (pilot)
84	20002	Categorical (multiple)	Non-cancer illness code, self-reported
85	20008	Continuous	Interpolated Year when non-cancer illness first diagnosed
86	20009	Continuous	Interpolated Age of participant when non-cancer illness first diagnosed
87	20015	Continuous	Sitting height
88	20041	Categorical (single)	Reason for skipping weight
89	20045	Categorical (single)	Reason for skipping waist
90	20046	Categorical (single)	Reason for skipping hip measurement
91	20047	Categorical (single)	Reason for skipping standing height
92	20048	Categorical (single)	Reason for skipping sitting height
93	20074	Integer	Home location at assessment - east co-ordinate (rounded)
94	20075	Integer	Home location at assessment - north co-ordinate (rounded)
95	20077	Integer	Number of diet questionnaires completed
96	20078	Date	When diet questionnaire completion requested
97	20079	Categorical (single)	Day-of-week questionnaire completion requested
98	20080	Categorical (single)	Day-of-week questionnaire completed
99	20081	Integer	Hour-of-day questionnaire completed
100	20082	Integer	Duration of questionnaire
101	20083	Integer	Delay between questionnaire request and completion
102	20085	Categorical (multiple)	Reason for not eating or drinking normally
103	20086	Categorical (multiple)	Type of special diet followed
104	20095	Categorical (multiple)	Size of white wine glass drunk
105	20096	Categorical (multiple)	Size of red wine glass drunk
106	20097	Categorical (multiple)	Size of rose wine glass drunk
107	20116	Categorical (single)	Smoking status
108	20117	Categorical (single)	Alcohol drinker status
109	20160	Categorical (single)	Ever smoked
110	20161	Continuous	Pack years of smoking
111	20162	Continuous	Pack years adult smoking as proportion of life span exposed to smoking
112	20401	Categorical (single)	Ever addicted to any substance or behaviour
113	20403	Categorical (single)	Amount of alcohol drunk on a typical drinking day
114	20404	Categorical (single)	Ever physically dependent on alcohol
115	20405	Categorical (single)	Ever had known person concerned about, or recommend reduction of, alcohol consumption
116	20406	Categorical (single)	Ever addicted to alcohol
117	20407	Categorical (single)	Frequency of failure to fulfil normal expectations due to drinking alcohol in last year
118	20408	Categorical (single)	Frequency of memory loss due to drinking alcohol in last year
119	20409	Categorical (single)	Frequency of feeling guilt or remorse after drinking alcohol in last year
120	20410	Integer	Age when known person last commented about drinking habits
121	20411	Categorical (single)	Ever been injured or injured someone else through drinking alcohol
122	20412	Categorical (single)	Frequency of needing morning drink of alcohol after heavy drinking session in last year
123	20413	Categorical (single)	Frequency of inability to cease drinking in last year

Distinct Data Field	Unique Data Identifier*	Type	Description
124	20414	Categorical (single)	Frequency of drinking alcohol
125	20415	Categorical (single)	Ongoing addiction to alcohol
126	20416	Categorical (single)	Frequency of consuming six or more units of alcohol
127	20431	Categorical (single)	Ever addicted to a behaviour or miscellanous
128	20432	Categorical (single)	Ongoing behavioural or miscellanous addiction
129	20453	Categorical (single)	Ever taken cannabis
130	20454	Categorical (single)	Maximum frequency of taking cannabis
131	20455	Integer	Age when last took cannabis
132	20456	Categorical (single)	Ever addicted to illicit or recreational drugs
133	20457	Categorical (single)	Ongoing addiction or dependence on illicit or recreational drugs
134	20488	Categorical (single)	Physically abused by family as a child
135	20490	Categorical (single)	Sexually molested as a child
136	20503	Categorical (single)	Ever addicted to prescription or over-the-counter medication
137	20504	Categorical (single)	Ongoing addiction or dependence to over-the-counter medication
138	20544	Categorical (multiple)	Mental health problems ever diagnosed by a professional
139	20551	Categorical (multiple)	Substance of prescription or over-the-counter medication addiction
140	20552	Categorical (multiple)	Behavioural and miscellaneous addictions
141	21000	Categorical (single)	Ethnic background
142	21001	Continuous	Body mass index (BMI)
143	21002	Continuous	Weight
144	21003	Integer	Age when attended assessment centre
145	21022	Integer	Age at recruitment
146	22000	Categorical (single)	Genotype measurement batch
147	22001	Categorical (single)	Genetic sex
148	22002	Text	CEL files
149	22003	Continuous	Heterozygosity
150	22004	Continuous	Heterozygosity, PCA corrected
151	22005	Continuous	Missingness
152	22006	Categorical (single)	Genetic ethnic grouping
153	22007	Text	Genotype measurement plate
154	22008	Text	Genotype measurement well
155	22009	Continuous	Genetic principal components
156	22019	Categorical (single)	Sex chromosome aneuploidy
157	22020	Categorical (single)	Used in genetic principal components
158	22021	Categorical (single)	Genetic kinship to other participants
159	22022	Continuous	Sex inference X probe-intensity
160	22023	Continuous	Sex inference Y probe-intensity
161	22024	Continuous	DNA concentration
162	22025	Continuous	Affymetrix quality control metric "Cluster.CR"
163	22026	Continuous	Affymetrix quality control metric "dQC"
164	22027	Categorical (single)	Outliers for heterozygosity or missing rate
165	22028	Categorical (single)	Use in phasing Chromosomes 1-22
166	22029	Categorical (single)	Use in phasing Chromosome X
167	22030	Categorical (single)	Use in phasing Chromosome XY
168	22100	Text	Chromosome XY genotype results
169	22101	Text	Chromosome 1 genotype results
170	22102	Text	Chromosome 2 genotype results
171	22103	Text	Chromosome 3 genotype results
172	22104	Text	Chromosome 4 genotype results
173	22105	Text	Chromosome 5 genotype results
174	22106	Text	Chromosome 6 genotype results
175	22107	Text	Chromosome 7 genotype results
176	22108	Text	Chromosome 8 genotype results
177	22109	Text	Chromosome 9 genotype results
178	22110	Text	Chromosome 10 genotype results
179	22111	Text	Chromosome 11 genotype results
180	22112	Text	Chromosome 12 genotype results
181	22113	Text	Chromosome 13 genotype results
182	22114	Text	Chromosome 14 genotype results
183	22115	Text	Chromosome 15 genotype results
184	22116	Text	Chromosome 16 genotype results
185	22117	Text	Chromosome 17 genotype results
186	22118	Text	Chromosome 18 genotype results
187	22119	Text	Chromosome 19 genotype results
188	22120	Text	Chromosome 20 genotype results
189	22121	Text	Chromosome 21 genotype results

Distinct Data Field	Unique Data Identifier*	Type	Description
190	22122	Text	Chromosome 22 genotype results
191	22123	Text	Chromosome X genotype results
192	22124	Text	Chromosome Y genotype results
193	22125	Text	Mitochondrial genotype results
194	22182	Curve	HLA imputation values
195	22700	Date	Date first recorded at location
196	22702	Integer	Home location - east co-ordinate (rounded)
197	22704	Integer	Home location - north co-ordinate (rounded)
198	22800	Text	Chromosome XY imputation and haplotype results
199	22801	Text	Chromosome 1 imputation and haplotype results
200	22802	Text	Chromosome 2 imputation and haplotype results
201	22803	Text	Chromosome 3 imputation and haplotype results
202	22804	Text	Chromosome 4 imputation and haplotype results
203	22805	Text	Chromosome 5 imputation and haplotype results
204	22806	Text	Chromosome 6 imputation and haplotype results
205	22807	Text	Chromosome 7 imputation and haplotype results
206	22808	Text	Chromosome 8 imputation and haplotype results
207	22809	Text	Chromosome 9 imputation and haplotype results
208	22810	Text	Chromosome 10 imputation and haplotype results
209	22811	Text	Chromosome 11 imputation and haplotype results
210	22812	Text	Chromosome 12 imputation and haplotype results
211	22813	Text	Chromosome 13 imputation and haplotype results
212	22814	Text	Chromosome 14 imputation and haplotype results
213	22815	Text	Chromosome 15 imputation and haplotype results
214	22816	Text	Chromosome 16 imputation and haplotype results
215	22817	Text	Chromosome 17 imputation and haplotype results
216	22818	Text	Chromosome 18 imputation and haplotype results
217	22819	Text	Chromosome 19 imputation and haplotype results
218	22820	Text	Chromosome 20 imputation and haplotype results
219	22821	Text	Chromosome 21 imputation and haplotype results
220	22822	Text	Chromosome 22 imputation and haplotype results
221	22823	Text	Chromosome X imputation and haplotype results
222	23098	Continuous	Weight
223	23099	Continuous	Body fat percentage
224	23100	Continuous	Whole body fat mass
225	23101	Continuous	Whole body fat-free mass
226	23102	Continuous	Whole body water mass
227	23104	Continuous	Body mass index (BMI)
228	23105	Continuous	Basal metabolic rate
229	23106	Continuous	Impedance of whole body
230	23127	Continuous	Trunk fat percentage
231	23128	Continuous	Trunk fat mass
232	23129	Continuous	Trunk fat-free mass
233	23130	Continuous	Trunk predicted mass
234	40000	Date	Date of death
235	40001	Categorical (single)	Underlying (primary) cause of death: ICD10
236	40002	Categorical (single)	Contributory (secondary) causes of death: ICD10
237	40018	Categorical (single)	Death report format
238	41078	Integer	Episodes containing "Diagnoses - secondary ICD10" data
239	41080	Integer	Episodes containing "Operative procedures - secondary OPCS4" data
240	41082	Integer	Episodes containing "Dates of operations" data
241	41083	Integer	Episodes containing "Episode start date" data
242	41084	Integer	Episodes containing "Episode end date" data
243	41096	Integer	Episodes containing "Date of admission to hospital" data
244	41101	Integer	Episodes containing "Date of discharge from hospital" data
245	41142	Integer	Episodes containing "Diagnoses - main ICD10" data
246	41146	Integer	Episodes containing "Operative procedure - main OPCS4" data
247	41148	Integer	Episodes containing "Date of operation" data
248	41252	Integer	Episodes containing "Inpatient record format" data
249	100001	Continuous	Food weight
250	100002	Continuous	Energy
251	100003	Continuous	Protein
252	100004	Continuous	Fat
253	100005	Continuous	Carbohydrate
254	100006	Continuous	Saturated fat
255	100007	Continuous	Polyunsaturated fat

Distinct Data Field	Unique Data Identifier*	Type	Description
256	100008	Continuous	Total sugars
257	100009	Continuous	Englyst dietary fibre
258	100010	Categorical (single)	Portion size
259	100020	Categorical (single)	Typical diet yesterday
260	100022	Continuous	Alcohol
261	100023	Continuous	Starch
262	100026	Categorical (single)	Daily dietary data credible
263	100240	Categorical (single)	Coffee consumed
264	100250	Categorical (single)	Instant coffee intake
265	100270	Categorical (single)	Filtered coffee intake
266	100290	Categorical (single)	Cappuccino intake
267	100300	Categorical (single)	Latte intake
268	100310	Categorical (single)	Espresso intake
269	100330	Categorical (single)	Other coffee type
270	100360	Categorical (single)	Decaffeinated coffee
271	100390	Categorical (single)	Tea consumed
272	100400	Categorical (single)	Standard tea intake
273	100410	Categorical (single)	Rooibos tea intake
274	100420	Categorical (single)	Green tea intake
275	100430	Categorical (single)	Herbal tea intake
276	100440	Categorical (single)	Other tea intake
277	100470	Categorical (single)	Decaffeinated tea
278	100580	Categorical (single)	Alcohol consumed
279	100590	Categorical (single)	Red wine intake
280	100630	Categorical (single)	Rose wine intake
281	100670	Categorical (single)	White wine intake
282	100710	Categorical (single)	Beer/cider intake
283	100720	Categorical (single)	Fortified wine intake
284	100730	Categorical (single)	Spirits intake
285	100740	Categorical (single)	Other alcohol intake
286	105010	Time	When diet questionnaire completed
287	105030	Time	When diet questionnaire started

Appendix C. Geographic distribution of participants

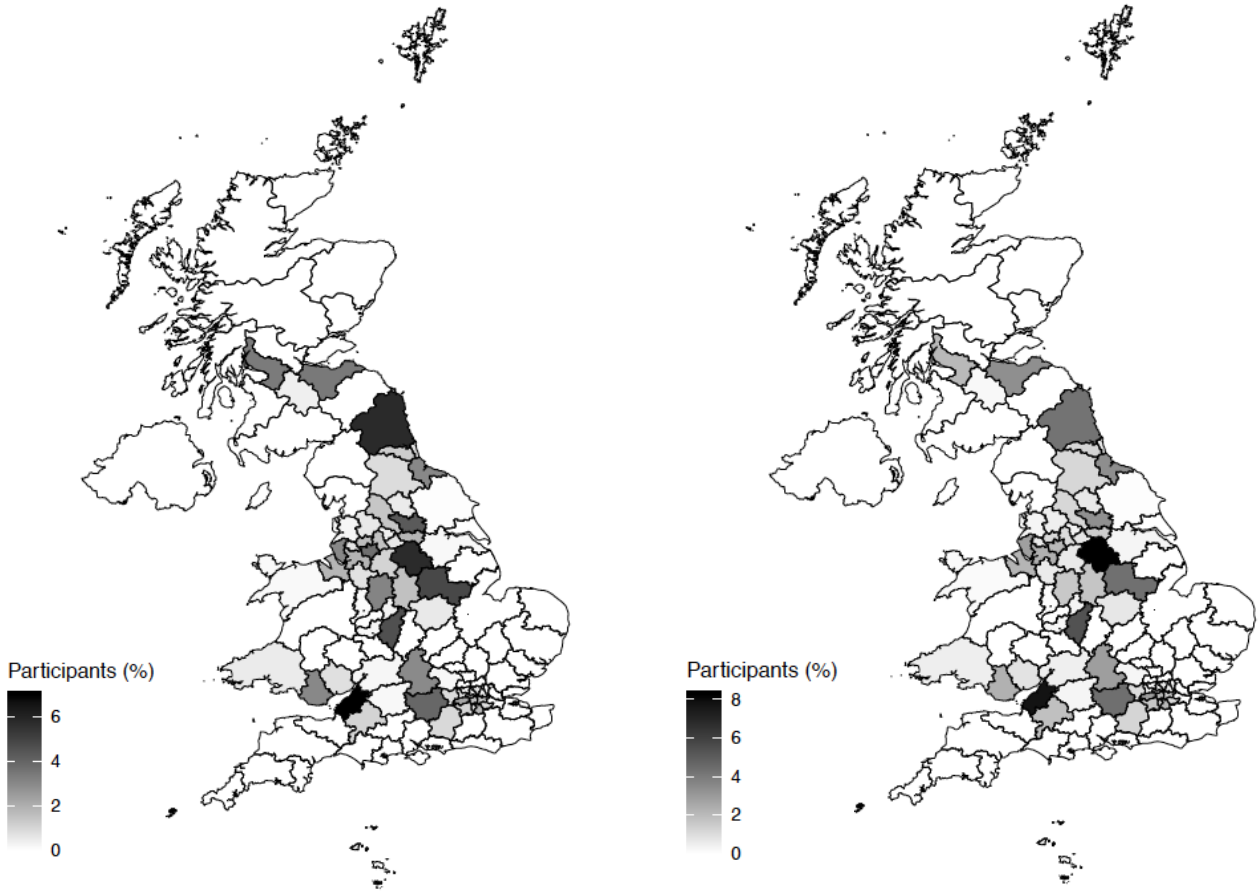


Figure C1. Maps showing the percentage of total participants from each geographic region in the UK Biobank cohort (left) and the study subsample (right).

Appendix D. Summary of training set phenotype and environment EDA results

Table D1. Performance of the addictions and smoking questionnaire sections

Questionnaire section	Addictions	Smoking
Parent category	Mental health	Lifestyle and environment
Category ID	141	100058
Number of distinct data fields in category	12	33
Participants who answered at least 1 question, no. (%)	44,886 (78.8)	56,994 (100)
Of the participants who answered at least 1 question, no. (%) participants who answered "Do not know," "Prefer not to answer," etc. for:		
0 questions	44,371 (98.9)	45,317 (79.5)
1 question	481 (1.1)	9,128 (16.0)
2 questions	30 (0.1)	1,994 (3.5)
3+ questions	4 (<0.1)	555 (1.0)

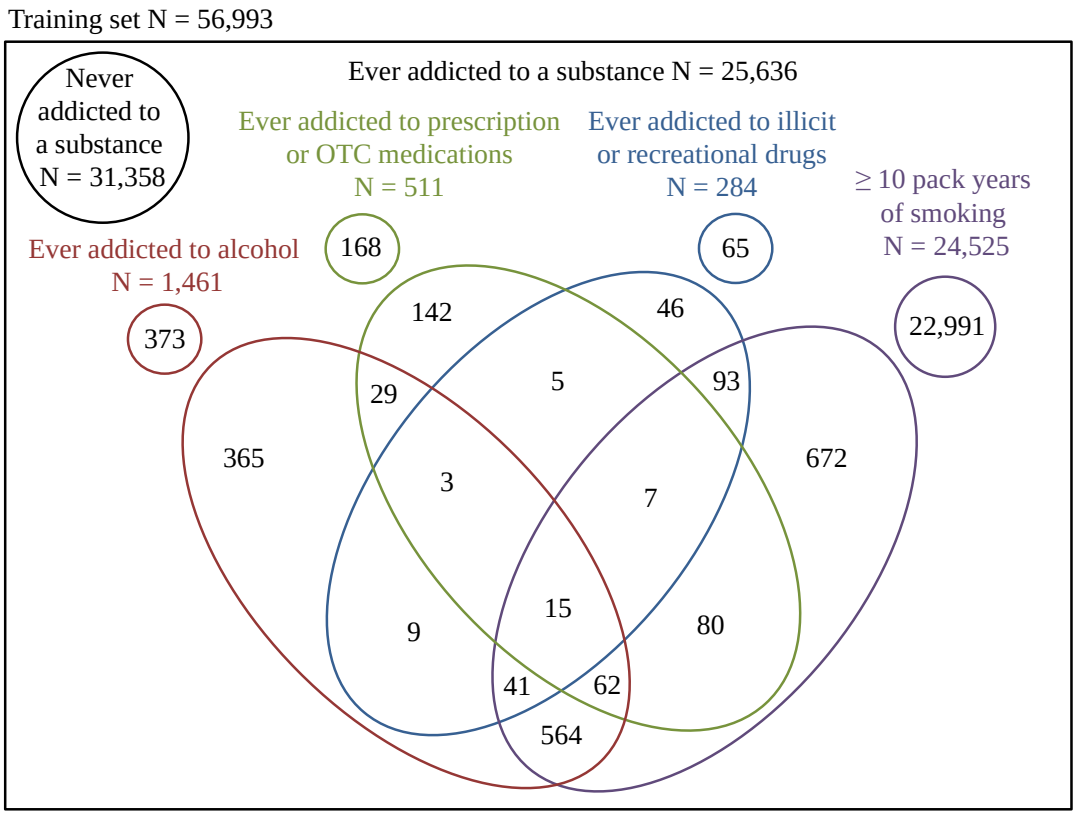


Figure D1. Venn diagram of participants by history of or current substance addiction(s). Colored circles contain participants with missing information for at least 1 of the 3 other substance addictions.

Table D2. Study outcomes by number of 24-hour dietary recall questionnaires completed

	Training set		UK Biobank*		
Number of 24-hour dietary recall questionnaires completed	1 N = 18,180 (31.9%)	2+ N = 38,814 (68.1%)	0 N = 91,779 (48.6%)	1 N = 31,188 (16.5%)	2+ N = 65,801 (34.9%)
Obesity, no. (%)					
BMI ≥ 30 kg/m^2	4,707 (25.9)	7,988 (20.6)	26,582 (29.0)	8,079 (25.9)	13,573 (20.6)
BMI < 30 kg/m^2	13,473 (74.1)	30,826 (79.4)	65,197 (71.0)	23,109 (74.1)	52,228 (79.4)
Substance addiction, no. (%)					
Substance addiction	10,491 (57.7)	15,145 (39.0)	71,625 (78.0)	18,209 (58.4)	26,090 (39.6)
No substance addiction	7,689 (42.3)	23,669 (61.0)	20,154 (22.0)	12,979 (41.6)	39,711 (60.4)

*Provided for comparison. Includes UK Biobank participants of all ancestries with a status for both obesity and substance addiction.

Table D3. Estimated caffeine content per coffee or tea drink

UK Biobank Unique Data Identifier	UK Biobank Description	UK Biobank Quantity	USDA* Quantity	USDA* Caffeine (mg)	Caffeine (mg) per UK Biobank Quantity
100270	Filtered coffee intake	1 cup/mug	1 cup (8 fl. oz.)	95	95
100310	Espresso intake	1 cup	1 fl. oz.	64	64
100290	Cappuccino intake	1 cup/mug	—	—	64
100300	Latte intake	1 cup/mug	—	—	64
100250	Instant coffee intake	1 cup/mug	1 tsp.	31	31
100330	Other coffee type	1 cup/mug	—	—	64
100400	Standard tea intake	1 cup/mug	1 cup (8 fl. oz.) “Black Tea”	47	47
100420	Green tea intake	1 cup/mug	1 cup (8 fl. oz.) “Tea”	26	26
100440	Other tea intake	1 cup/mug	—	—	26

*U.S. Department of Agriculture

Table D4. Spearman’s correlation coefficient matrix of reward system phenotypes

	Total fat (%)	Total sugars (%)	Caffeine (mg)	Alcohol (g)	Cigarettes (no.)
Total fat (%)	1.00	-0.42	0.05	-0.21	0.06
Total sugars (%)	—	1.00	-0.04	-0.31	-0.05
Caffeine (mg)	—	—	1.00	0.05	0.06
Alcohol (g)	—	—	—	1.00	0.00
Cigarettes (no.)	—	—	—	—	1.00

Table D5. Study outcomes by reporting of physical or sexual abuse as a child

	Training set	Physical or sexual abuse as a child reported	Physical or sexual abuse as a child NOT reported
	N = 56,994 (100%)	N = 44,520 (78.1%)	N = 12,474 (21.9%)
Obesity, number (%)			
BMI ≥ 30 kg/m^2	12,695 (22.3)	8,941 (20.1)	3,754 (30.1)
BMI < 30 kg/m^2	44,299 (77.7)	35,579 (79.9)	8,720 (69.9)
Substance addiction, number (%)			
Substance addiction	25,636 (45.0)	13,409 (30.1)	12,227 (98.0)
No substance addiction	31,358 (55.0)	31,111 (69.9)	247 (2.0)

Appendix E. Summary of phenotype score development results

Table E1. Reward system phenotype distribution fit statistics

Reward system phenotype	One-sample Kolmogorov–Smirnov test statistic (D)*		
	Normal distribution	Poisson distribution (λ=1)	Log(*x*+1)-normal distribution
Nutrient intake			
Total fat, mean % energy yesterday	0.9998	—	0.9930
Total sugars, mean % energy yesterday	0.9992	—	0.9777
Drug use			
Caffeine, mean yesterday (mg)	0.9363	0.9361	0.9294
Alcohol, mean yesterday (g)	0.6342	0.6139	0.5661
Cigarettes, mean no. smoked daily	0.5000	0.5399	0.5000

*$P < 0.05$ for all phenotypes and distributions.

Table E2. Drug use phenotype distribution fit statistics

Drug use phenotype	Poissoness plot test statistic ($\lvert\lambda_{ML} - \exp(\text{slope})\rvert$)	
	Participants with zero drug use included	Participants with zero drug use excluded
Caffeine, mean yesterday (mg)	186.822	1.825
Alcohol,		
mean yesterday (g)	20.119	20.110
mean daily (g)*	35.577	35.757
Cigarettes,		
mean no. smoked daily	4.538	5.071
with 1 outlier removed	5.247	4.801

*Provided for comparison. Derived from (1) the alcohol section (category ID 100051) of the UK Biobank lifestyle and environment questionnaire and (2) alcohol units per drink as defined by the UK National Health Service.

Table E3. Pearson correlation coefficient matrix of power-transformed drug use phenotypes and nutrient intake phenotypes

	Total fat (%)	Total sugars (%)	Caffeine (mg)	Alcohol (g)	Cigarettes (no.)
Total fat (%)	1.00	-0.45	0.05	-0.22	0.06
Total sugars (%)	—	1.00	-0.05	-0.31	-0.05
Caffeine (mg)	—	—	1.00	0.07	0.07
Alcohol (g)	—	—	—	1.00	0.00
Cigarettes (no.)	—	—	—	—	1.00

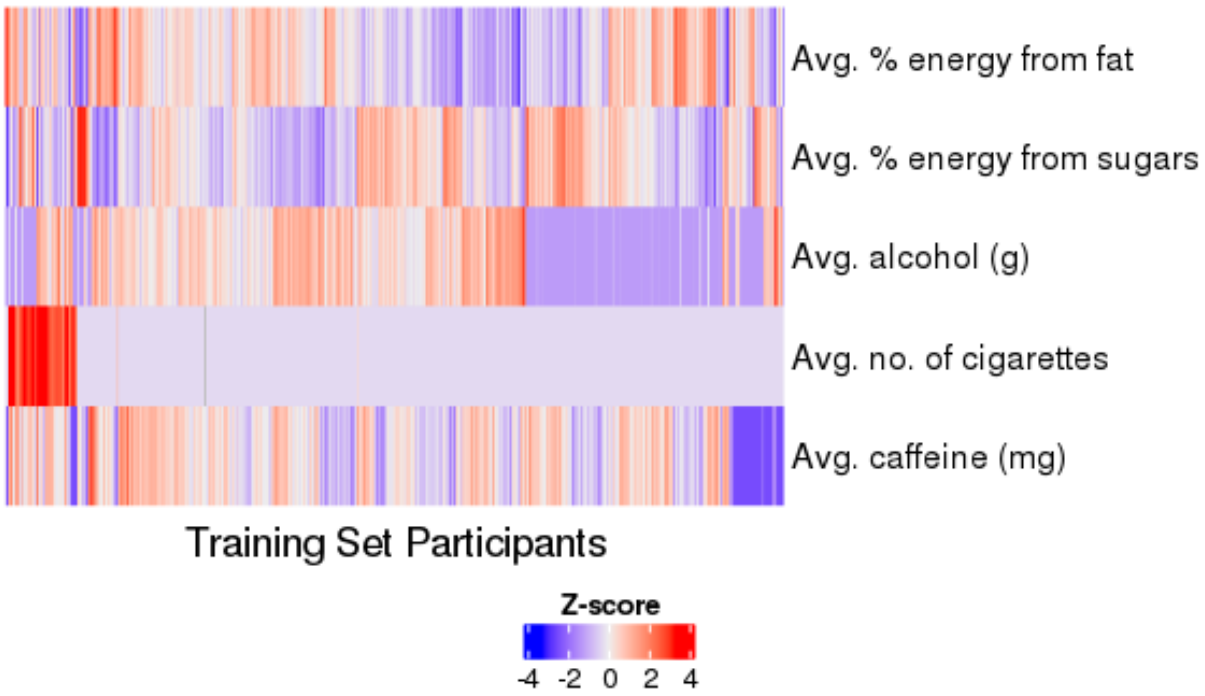


Figure E1. Average-linkage hierarchical clustering of 10K training set participants chosen at random using the Euclidean distance between their power-transformed drug use and nutrient intake phenotypes. Average number of cigarettes is marked in grey for participants who did not report their current daily use at least once (N = 298, 0.5% of the training set).

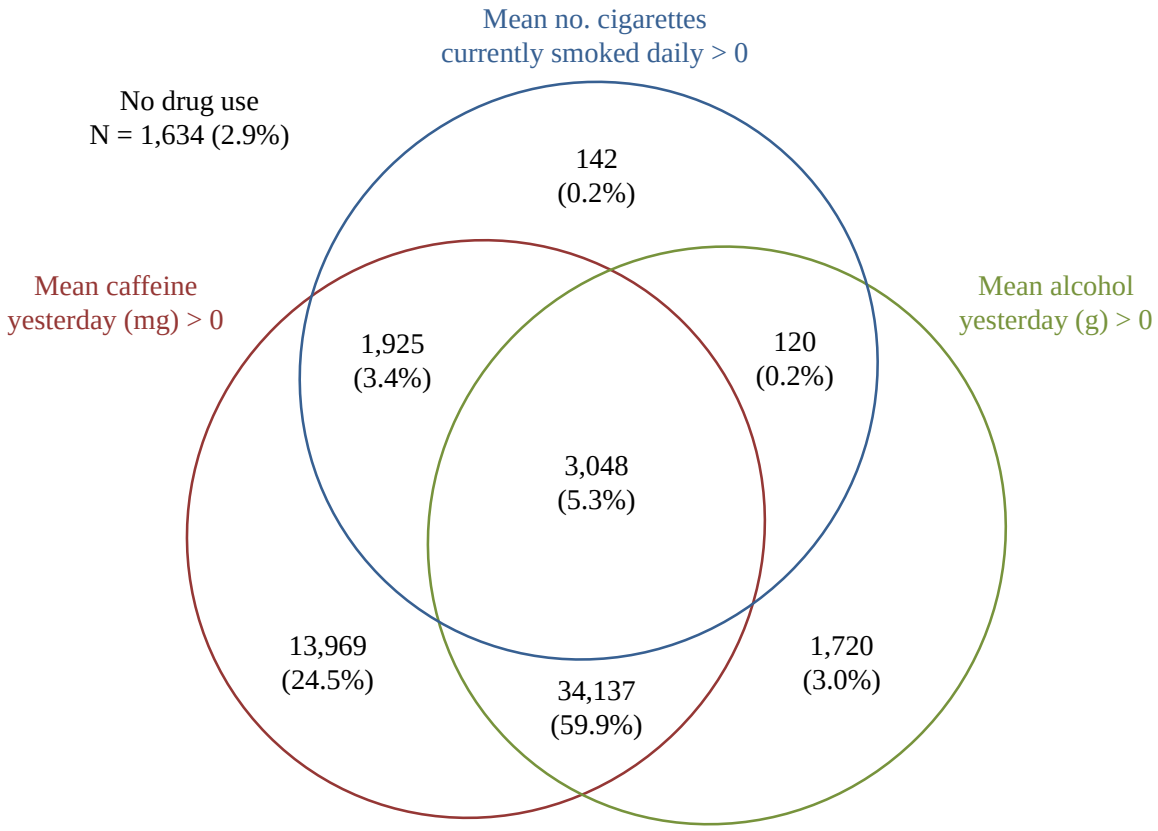


Figure E2. Venn diagram of participants by drug use phenotype(s). Participants who did not report the number of cigarettes currently smoked daily at least once are not shown (N = 298, 0.5% of the training set).

Table E4. UK Biobank data fields comprising the null phenotype scores

	Distinct Data Field	Unique Data Identifier (UDI)	Description
1	4	34	Year of birth
2	99	20081	Hour-of-day questionnaire completed
3	161	22024	DNA concentration
4	197	22704	Home location - north co-ordinate (rounded)
5	238	41078	Episodes containing "Diagnoses - secondary ICD10" data

Table E5. Spearman's correlation coefficients between null and reward system phenotypes

	Spearman's correlation coefficient				
Null Phenotype	Total fat (%)	Total sugars (%)	Caffeine (mg)	Alcohol (g)	Cigarettes (no.)
Year of birth	0.04	-0.06	-0.06	-0.04	0.1
Hour-of-day questionnaire completed	0	0.06	0	-0.06	-0.02
DNA concentration	0.01	-0.02	0	0.01	0.08
Home location - north co-ordinate (rounded)	-0.02	0	-0.02	-0.04	0
Episodes containing "Diagnoses - secondary ICD10" data	0	0	-0.03	-0.04	0.07

Appendix F. Summary of genotype and post-GWAS quality control results

Table F1. Genotype marker quality control results

Marker quality control criteria	UK Biobank (N = 488,377)	Training set (N = 56,993)
Genotyped and imputed markers (N = 97,059,328)		
Minor allele frequency ≤ 0.1% among participants	76,755,998 (79.1%)	79,854,667 (82.3%)
Information quality score[81] ≤ 0.3 among participants	16,258,690 (16.8%)	—
Passed marker quality control criteria	20,210,907 (20.8%)	—
Genotyped markers only (N = 805,426)		
Missing in ≥ 1% of participants	148,075 (18.4%)	146,007 (18.1%)
Minor allele frequency ≤ 0.1% among participants	75,832 (9.4%)	82,351 (10.2%)
Missing in ≥ 1% of participants, or minor allele frequency ≤ 0.1%	610,046 (75.7%)	605,623 (75.2%)

Table F2. Sample quality control training set results

Sample quality control criteria	UK Biobank (N = 488,377)	Training set (N = 56,993)
Outliers for heterozygosity or missing rate among UK Biobank participants*	968 (0.20%)	92 (0.16%)
Sex discordant	652 (0.13%)	40 (0.07%)
Genotypes not imputed across all chromosomes	935 (0.19%)	186 (0.33%)
Third-degree relative or closer to at least one other UK Biobank participant	147,731 (30.2%)	17,198 (30.2%)
Passed sample quality control criteria	—	39,710 (69.7%)

*Performed by the UK Biobank researchers on a set of 621,642 high quality SNPs and accounting for population stratification.

Table F3. Post marker quality control training set results

Marker quality control criteria	Training set (N = 39,710)
Genotyped and imputed markers (N = 20,271,549*)	
Minor allele frequency ≤ 0.1% among participants	3,425,000 (16.9%)
Genotyped markers only (N = 610,046)	
Missing in ≥ 1% of participants	1,718 (0.3%)
Minor allele frequency ≤ 0.1% among participants	5,671 (0.9%)

*Of 20,210,907 total genotyped and imputed markers that passed quality control, multi-allelic variants (0.3%) were counted as independent markers.

Table F4. Rare reward system aggregate phenotype-associated loci (MAF <0.0001)

Locus	Phenotype score	Chr	Locus boundaries	GWS variants at locus (no.)	Index variant	Reference/ minor alleles	UKB MAF	INFO score[81]	β	*q*-value	Closest gene to index variant
1	Max	2	37386144 - 37387207	3	rs10177345	C/G	0.001	0.92	3.44	0.018	*EIF2AK2*
2	Max	3	23529598 - 23532429	2	rs17013356	A/G	0.001	0.98	5.01	0.003	*UBE2E2*
3	Max	3	73167300	1	rs144259811	T/G	0.001	0.89	3.19	0.009	*PPP4R2*
4	Max	3	79727452 - 79746041	5	rs115208573	C/T	0.002	0.95	3.48	0.008	*ROBO1*
5	Max	3	97509046 - 97530462	2	chr3:97509046	TAAGAG/T	0.001	0.94	2.94	0.008	*ARL6*
6	Max	3	97841959	1	rs115858807	G/A	0.001	0.92	3.27	0.009	*OR5H1*
7	Max	3	100244947	1	rs9839721	A/G	0.003	0.97	1.97	0.012	*TMEM45A*
8	Max	4	16448564 - 16448572	2	rs75854427	T/C	0.001	0.85	2.80	<0.001	*LDB2*
9	Max	4	190166950 - 190189704	4	rs11941095	A/G	0.003	0.94	2.76	<0.001	*FRG1*
10	Both	14	21356821 - 21427751	2	rs10132681	G/A	0.001	0.89	5.35	0.013	*RNASE2*
11	Max	17	55699351	1	rs148796622	G/T	0.001	0.83	9.01	0.011	*MSI2*
12	Max	20	3643595	1	rs680606	C/T	0.002	0.88	3.95	0.015	*GFRA4*
13	Sum	20	59876307	1	rs73319112	G/T	0.002	0.95	8.90	0.007	*CDH4*

**Max, maximum; Chr., chromosome; GWS, genome-wide significant (*q*-value ≤ 0.025 for either phenotype score); MAF, minor allele frequency; UKB, UK Biobank. GWS variant with smallest *q*-value at locus is the index variant. Exact β and *q*-values for GWS variants with MAF < 0.0005 should be interpreted with caution. Reference assembly GRCh37/hg19.

Figure F1. Representative SNP cluster plots for rs2470893 from UK Biobank batches 1-4.[106]