HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2025 March 31.

Published in final edited form as:

Nat Genet. 2024 July ; 56(7): 1527–1536. doi:10.1038/s41588-024-01793-9.

Synthetic surrogates improve power for genome-wide association studies of partially missing phenotypes in population biobanks

Zachary R. McCaw[1,*], **Jianhui Gao**[2], **Xihong Lin**[1,3,†], **Jessica Gronsbell**[2,4,*,†]

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, US.

[2]Department of Statistical Sciences, University of Toronto, Toronto, Ontario, CA.

[3]Department of Statistics, Harvard University, Cambridge, MA, US.

[4]Department of Computer Science, University of Toronto, Toronto, Ontario, CA.

Abstract

Within population biobanks, incomplete measurement of certain traits limits power for genetic discovery. Machine learning (ML) is increasingly used to impute the missing values from the available data. However, performing genome-wide association studies (GWAS) on imputed traits can introduce spurious associations, identifying genetic variants not associated with the original trait. Here we introduce a new method, synthetic surrogate (SynSurr) analysis, which makes GWAS on imputed phenotypes robust to imputation errors. Rather than replacing missing values, SynSurr jointly analyzes the original and imputed traits. We show that SynSurr estimates the same genetic effect as standard GWAS, and improves power in proportion to the quality of the imputations. SynSurr requires a commonly-made missing at random assumption, but relaxes the requirements of existing imputation methods by not requiring correct model specification. We present extensive simulations and ablation analyses to validate SynSurr, and apply it to empower GWAS of dual-energy x-ray absorptiometry traits within the UK Biobank.

Introduction

The emergence of high-quality population biobanks, such as FinnGenn (> 200,000 individuals), the Million Veteran Program (> 500,000 veterans), and the UK Biobank (> 400,000 individuals), is empowering new genetic discoveries[1-3]. While these deeply phenotyped resources have created an unprecedented opportunity to increase both the scope and precision of genome-wide association studies (GWAS), they have simultaneously

*Correspondence to: zmccaw@alumni.harvard.edu and j.gronsbell@utoronto.ca.
†:Jointly supervised this work.

Author Contributions

ZRM, XL, and JeG designed the study and the experiments. ZRM implemented the software with input from XL. ZRM, JeG, and JiG performed the simulation experiments. JiG conducted analyses of the UK Biobank data. ZRM performed the overlap analysis. ZRM and JeG wrote the first draft of the manuscript and all co-authors provided intellectual revisions.

Competing interests

ZRM is currently an employee of insitro, but was not at the time of this work, and his employer had no role in this study. The remaining authors have no competing interests to declare.

introduced new statistical challenges[4,5]. Among these challenges, incompletely measured or partially missing phenotypes are a key issue[6]. Missingness often arises due to the difficulty, expense, or invasiveness of ascertaining the target phenotype[7]. Examples of partially missing phenotypes from the UK Biobank (UKBB)[8] include body composition phenotypes obtained from dual-energy x-ray absorptiometry (DEXA) scans[9], neurological[10] and cardiac[11] structural features extracted from functional magnetic resonance imaging, optic morphology parameters extracted from retinal fundus images[12] , and sleep wake patterns extracted from accelerometry trackers[13]. Each of these phenotypes was ascertained, at least initially, in only a subset of the cohort.

Restricting GWAS to only those individuals with observed phenotypes substantially diminishes power. As biobanks contain extensive demographic and clinical information, researchers can often impute missing phenotypes from the available data[14,15], and increasingly do so via machine learning[12,16-20]. Armed with an imputation model, the analysis can proceed in several ways. One approach, which we call *proxy GWAS*, predicts the target phenotype for all subjects then performs GWAS on the predicted values. Although proxy GWAS may perform well if the imputation model is highly accurate, this approach incurs bias and identifies false positives associations when the imputation model is inaccurate. Another approach, *single-imputation*, imputes the target phenotype for unlabeled subjects only, then performs GWAS on the combination of observed and imputed values. Single-imputation is statistically invalid because it ignores imputation uncertainty and treats the observed and imputed values as originating from the same distribution[21-24]. Multiple imputation[25-27] has been proposed to overcome the limitations of single imputation. In this approach, several sets of imputations are generated from a probabilistic model, analyzed in parallel, then combined via Rubin's rules, correctly accounting for imputation uncertainty. The major limitation of multiple imputation is the necessity for a correctly specified imputation model[27-30]. As we show here, if the imputation model is misspecified, as is most likely in practice, the estimates resulting from multiple imputation are biased.

To overcome the limitations of proxy- and imputation-based GWAS, we introduce synthetic surrogate ("SynSurr") analysis for GWAS of a partially missing target phenotype. Within a model-building dataset, we first train an imputation model for predicting the target phenotype on the basis of available data. Next, within the GWAS dataset, the imputation model is applied to generate an imputation or "synthetic surrogate" phenotype for all subjects. The partially-observed target phenotype and the synthetic surrogate are then *jointly* analyzed within a bivariate outcome framework[31]. We illustrate the methodological advantages and practical implementation of SynSurr through extensive analyses of simulated and real data. These analyses illustrate three key properties of SynSurr, which are backed by theoretical derivations provided in the Supplementary Methods. First, SynSurr is *robust* to imputation error, meaning that neither bias nor loss of power are incurred when the model that generates the synthetic surrogate is misspecified. Second, SynSurr is more *powerful* than standard GWAS, with the power advantage increasing with the proportion of target-missingness and the target-surrogate correlation. Third, SynSurr is always *valid*, meaning that the effect sizes are unbiased and the type I error is properly controlled, irrespective of the imputation model.

Our real-data analyses in the UK Biobank (UKBB) include an ablation analysis of two phenotypes with minimal missingness, height and forced expiratory volume in 1 second (FEV1), and an application of SynSurr to 6 body composition phenotypes, measured by DEXA, with substantial missingness. The ablation analysis demonstrates that SynSurr and standard GWAS identify all the same variants in the absence of missingness, but that SynSurr is uniformly more powerful in the presence of missingness. SynSurr achieves this power advantage not by inflating the false discovery rate or distorting the estimated genetic effect, but by leveraging the surrogate outcome to obtain more precise estimation (i.e., smaller standard errors). The application to the DEXA phenotypes demonstrates the substantial opportunity for improved power with SynSurr. Compared to standard GWAS, SynSurr identified on average 21.5× as many genome-wide significant variants, and did so at 3.3× the level of significance. Moreover, the variants identified by SynSurr are relevant to body composition, being significantly enriched for salient gene sets and overlapping substantially with previously reported associations from the GWAS Catalog[32].

Results

Overview of Method

Fig. 1 provides a graphical overview of SynSurr, which aims to empower GWAS of a partially missing target phenotype. The data are first split into model-building and inference sets. The model-building data should contain surrogate variables (i.e., the inputs to the imputation model) and at least some subjects with observed target phenotypes, assuming the imputation model is trained in a (semi-)supervised manner. The inference set, in which GWAS is performed, requires genetics and surrogates, and is assumed to be only partially labeled, meaning the target outcome is partially missing.

Within the model-building data, an imputation model is trained to predict the target phenotype from surrogate information. Predictions of the target phenotype are then generated for *all* subjects in the inference set (not only those with missing phenotypes). Because the imputation model combines multiple surrogates to predict the target phenotype, we refer to its output as a *synthetic surrogate*. Unlike standard imputation, the synthetic surrogate is maintained as a separate and distinct outcome from the target phenotype. Finally, within the inference data, GWAS is performed by jointly regressing the partially-missing target phenotype and the fully-observed synthetic surrogate on genotype and covariates.

The Methods section provides a mathematical description of the SynSurr model and an overview of the estimation procedure. The Supplementary Methods provide a detailed derivation of maximum likelihood estimates for all model parameters, their standard errors, and a Wald test for evaluating the association between genotype and the target phenotype.

Generation of synthetic surrogates

SynSurr depends on the availability of a synthetic surrogate $\hat{Y}$ which is predictive of the target phenotype Y. We focus on the setting where the surrogate is a prediction of Y from an ML model. This setting is particularly relevant in population biobanks where certain

phenotypes are too invasive, expensive, or time-consuming to measure for the entire cohort, but can be predicted from a model trained on available surrogate data (e.g., information from electronic health records or baseline assessments). The inputs to the imputation model (i.e., the model that generates the synthetic surrogate) should not include genetics G, and may or may not include covariates X adjusted for during the GWAS. In training the imputation model, the goal is to obtain a prediction $\hat{Y}$ that is highly correlated with the target phenotype Y *after adjusting for any covariates included in the GWAS*. As detailed in the Supplementary Methods section, a stronger residual correlation leads to increased power. To capture the potentially complex relationship between a set of covariates and the phenotype of interest, we recommend generating $\hat{Y}$ from a nonlinear model. For tabular data settings with well-defined covariates, tree-based models such as random forest[33] or Extreme Gradient Boosting (XGBoost)[34] generally perform well and are straightforward to train. Neural network may be particularly advantageous for models using images or free-text.

SynSurr overcomes the pitfalls of imputation-based inference

Recall that our goal is to perform inference on the association between genotype G and a partially-missing target phenotype Y. As the synthetic surrogate $\hat{Y}$ is available for all subjects, one option is to perform proxy GWAS on $\hat{Y}$ in place of Y. This approach, however, changes the research question from studying the association between Y and G (i.e., β_G in Equation 4) to studying that between $\hat{Y}$ and G (i.e., α_G in Equation 4). Moreover, analyzing $\hat{Y}$ instead of Y can lead to spurious association if $\hat{Y}$ is an imperfect proxy for Y (see the ablation studies). To preserve the original research question, another common approach is to generate a completed outcome Y^*, where missing values of Y are replaced by $\hat{Y}$:

$$Y^* = \begin{cases} Y \text{ if } Y \text{ is observed,} \\ \hat{Y} \text{ if } Y \text{ is missing.} \end{cases}$$

A single-imputation analysis would perform GWAS on Y^* once, whereas multiple imputation performs several imputations of the missing data then combines the results via Rubin's rules[21,27]. As shown below, single-imputation leads to underestimation of the standard errors while multiple-imputation requires the imputation model to be correctly specified. In contrast, SynSurr jointly models $\hat{Y}$ with Y, allowing for missing data (Fig. 1). A key contribution of SynSyrr, with ramifications beyond GWAS, is to provide a framework for utilizing $\hat{Y}$ to improve inference on Y *without* requiring that $\hat{Y}$ be generated from a correctly specified model.

As an example of the pitfalls that can arise when performing imputation-based inference, we generated phenotypes from the following model

$$Y = G\beta_G + X\beta_X + \epsilon. \quad (1)$$

where X is a covariate and ϵ a residual. The genetic effect is $\beta_G = 0.1$, corresponding to a variant with $h^2 = 1\,\%$. The total sample size is $n = 10^4$ and 25% of the values of Y are missing. We compare the performance of an *oracle estimator*, which has access to Y before the introduction of missingness, the *standard estimator*, which only has access to the observed values of Y, single-imputation, multiple-imputation, and SynSurr. The imputation models are as follows:

1. A correctly specified model, which imputes Y using G and X.
2. A misspecified model, which imputes Y using G only.
3. A misspecified model, which imputes Y using X only.

Each imputation model was fit on an independent model-building data set of size 10^3 then applied to generate $\hat{Y}$ for all subjects in the inference data set. All estimators utilized Equation (1) as the association model, allowing for direct comparison of the estimates of β_G.

Fig. 2 and Supplementary Table 1 present the point estimates and standard errors (SEs) of the imputation-based estimators compared to SynSurr. The oracle estimator (green) is unbiased for β_G, and as a correctly specified maximum likelihood estimator, its SE is the best possible in the absence of missing data[35]. The standard estimator (orange) is also unbiased, and because observations with missing outcomes do not contribute to this estimator, its SE is larger than the oracle's. The single- (red) and multiple-imputation (blue) estimators are unbiased only when the imputation model is correctly specified. While the 95% confidence interval (CI) for the multiple-imputation estimator has proper coverage, as indicated by the agreement between the analytical (dotted) and empirical CIs (solid), the analytical CI for the single-imputation estimator falls short of the empirical CI. This occurs because the SE of the single-imputation estimator is underestimated. Consequently, inference based on the single-imputation estimator leads to an overstatement of significance (i.e., inflated type I error)[36]. Estimates based on an incorrectly specified imputation model are biased, whether the the misspecification was due to omission of the variable of interest (i.e., G) or a covariate (i.e., X)[26]. In contrast, SynSurr (purple) is unbiased regardless of what collection of covariates is used to generate $\hat{Y}$.

SynSurr is robust to the choice of surrogate

Unlike imputation-based inference, which is sensitive to correct specification of the imputation model, SynSurr is robust to the choice of synthetic surrogate in that it (i) consistently estimates the effect of G on Y regardless of the correlation between Y and $\hat{Y}$ and (ii) provides improved power over standard GWAS when the synthetic surrogate $\hat{Y}$ is correlated with the target phenotype Y. To demonstrate these points, we again simulated phenotypes from Equation (1) and varied the proportion of subjects with missing phenotypes and the correlation between Y and $\hat{Y}$ (see Supplementary Section 3.1). Extended Data Fig. 1 presents box plots of $\hat{\beta}_G - \beta_G$ for the standard GWAS estimator and the SynSurr estimator under different levels of missingness in Y. In panel A, the synthetic surrogate $\hat{Y}$ is completely uncorrelated with, and in fact independent of, the phenotype of interest Y. Nevertheless, SynSurr is unbiased and no less efficient than standard GWAS, as indicated

by consistent widths of the box plots. This demonstrates that SynSurr is robust to having an uninformative $\hat{Y}$. However, as demonstrated in panel B, when $\hat{Y}$ is correlated with Y, the precision of SynSurr increases with the extent of missingness (also see Supplementary Fig. 2). These findings are supported by a quantitative comparison of standard errors in Supplementary Table 2 as well as theoretical analysis (see Supplementary Section 1.4).

SynSurr controls type I error and improves power

Building on the findings of the previous section, we again simulated phenotypes from Equation (1) to evaluate the type I error and power of SynSurr across various missing rates, synthetic surrogates, and levels of SNP heritability. Type I error is the probability of incorrectly rejecting $H_0 : \beta_G = 0$ when $\beta_G = 0$ and power is the probability of correctly rejecting H_0 when $\beta_G \neq 0$. Fig. 3 demonstrates proper type I error control in that the SynSurr p-values are uniformly distributed under the null hypothesis across all simulation settings. Extended Data Table 1 presents the type I error, as well as power and the average χ^2 statistics at a SNP heritability of $h^2 = 0.5\,\%$ ($\beta_G \approx 0.07$). The type I error is consistently controlled, and power increases with both target missingness and the correlation of the synthetic surrogate. For instance, when the missingness rate is 90% and the correlation between the synthetic surrogate and the target phenotype is 0.75, there is a 27% increase in power relative to standard GWAS. Fig. 4 illustrates the benefit of SynSurr with respect to power across SNP heritabilities ranging from 0.1% to 1.0%. As the proportion of subjects with missing target phenotypes increases, the benefit of SynSurr with a well-correlated surrogate phenotype is increasingly apparent. Interestingly, the relative efficiency of SynSurr does not depend on the SNP heritability (Supplementary Fig. 3).

Evaluation on UK Biobank (UKBB) data

We next demonstrate the advantages of SynSurr over standard GWAS through multiple analyses in the UKBB[37]. Our first evaluation compares SynSurr and standard GWAS of height and FEV1 – two traits measured for nearly-all participants – as the target phenotype is increasingly ablated, providing a scenario in which the ground truth is known. We then perform SynSurr analysis of 6 incompletely measured DEXA traits. Bioelectrical impedance, an imprecise measure of body-composition, was recorded for most participants at baseline[37,38]. Ascertainment of DEXA scans, a highly-precise measure of body-composition, began with a pilot study of 5K randomly selected participants in 2014 and remains ongoing[39,40]. At the time of our study, DEXA traits were available for 30K participants, whereas impedance measurements were available for 500K participants, providing a natural opportunity for deploying SynSurr.

SynSurr outperforms standard GWAS with increasing ablation—Details of the ablation study are described in the Methods. Briefly, sets of kin were identified, then one subject was allocated to the GWAS data set while the remaining were allocated to the model-building data set. The synthetic surrogate was generated from a random forest trained on the model-building data. We note that having related subjects in the model-building and GWAS data sets is not problematic because 1. the GWAS data are not being used to evaluate generalization performance, and 2. the subjects within the GWAS data are independent.

For each phenotype, we performed an oracle GWAS prior to the introduction of missingness, which establishes the number of genome-wide significant (GWS) associations that would be detected if the target phenotype were fully observed. Then, between 25% and 90% of the target phenotypes were randomly ablated. Both standard and SynSurr GWAS were performed on the remaining data. Prior to the introduction of missingness, the correlation between the target and the synthetic surrogate in the GWAS data set was $R^2 = 0.67$ for height and $R^2 = 0.51$ for FEV1. Scatter plots of predicted vs. observed height and FEV1 in the model-building and GWAS data are shown in Supplementary Figs. 8-9. Table 1 presents the numbers of oracle associations recovered by both the standard and SynSurr GWAS. SynSurr consistently recovers a higher proportion of the oracle associations than standard. Importantly, as demonstrated in Supplementary Table 7, SynSurr does not achieve this higher recovery by having a higher false discovery rate (FDR). Rather, SynSurr is leveraging the correlated surrogate outcome to obtain more precise SEs (Supplementary Tables 8-9). Extended Data Fig. 2 verifies that SynSurr is estimating the same genetic effect as the oracle GWAS (as is standard GWAS, see Supplementary Fig. 11). Even with 90% of target phenotypes ablated, the R^2 for the genetic effects between SynSurr and oracle is 0.90 for height and 0.87 for FEV1. With 50% of target phenotypes ablated, the R^2 rises to 0.99 and 0.98 respectively, and in the absence of missingness, $R^2 = 1.00$.

Working within the ablation framework, we also compared SynSurr with imputation-based GWAS and Multi-Trait Analysis of GWAS (MTAG)[41]. Beginning with imputation, we focused on the setting of 50% missingness, comparing SynSurr with single- and multiple-imputation when the surrogates and imputations were either high quality (generated by random forest), low quality (generated by linear regression), permuted, or negated. The results are presented in Supplementary Tables 10-14. As expected, single-imputation fails to consistently control the FDR, and although multiple-imputation performed better in this regard, it was underpowered, identifying fewer than 20% associations then SynSurr. While permutation or negation compromised imputation-based inference, SynSurr was robust to permutation, performing comparably to standard GWAS, and invariant to negation. Extended Data Fig. 3 (height) and Supplementary Fig. 13 (FEV1) demonstrate that even when single- or multiple-imputation properly control the FDR, the estimated genetic effects are generally biased, whereas those of SynSurr are always unbiased.

For comparison with MTAG, we performed proxy GWAS (i.e., standard GWAS of $\hat{Y}$) then combined standard GWAS of Y with proxy GWAS via MTAG. The results are presented in Extended Data Table 2 (height) and Supplementary Table 18 (FEV1). Due to the imperfect correlation between Y and $\hat{Y}$, proxy GWAS was biased (Supplementary Fig. 14) and poorly controlled the FDR, identifying numerous significant associations not detected by the oracle. As a result, MTAG inherited an inflated FDR. For example, in the case of FEV1, proxy GWAS had a FDR of 86% while MTAG had a FDR rising from 19% in the absence of missingness to 67% at 90% missingness. In cases where MTAG did control the FDR, SynSurr generally provided higher power.

SynSurr empowers body composition GWAS—We next performed SynSurr analysis of 6 DEXA traits, following the same sample splitting procedure used for the ablation

analysis (see Methods). Among subjects in the GWAS data set, the average R^2 between the target phenotype and the synthetic surrogate was 0.80 (Extended Data Fig. 4). The distributions of covariates (Supplementary Fig. 16) and of predicted DEXA measurements (Extended Data Fig. 5) were similar between subjects with and without DEXA scans.

Figure 5 presents the number of GWS associations ($p < 5 \times 10^{-8}$) for standard GWAS and SynSurr GWAS, as well as the average χ^2 statistic at the union of variants that reached significance in either GWAS. A larger χ^2 statistic indicates higher power to detect an association. Standard GWAS identified between 8 and 10 GWS variants (8.3 on average), while SynSurr GWAS identified between 65 and 270 GWS variants, for an average of 179.5 (21.5-fold improvement; Extended Data Table 3). The average χ^2 statistic at GWS variants was 46.2 for SynSurr GWAS, compared with 14.1 for standard GWAS, a 3.3-fold improvement. To check control of the type I error, the SynSurr analysis was repeated using permuted phenotypes. The uniform quantile-quantile plots in Supplementary Fig. 18 show no evidence of type I error inflation under the null.

The Miami plots in Supplementary Fig. 19 indicate SynSurr can elevate a subthreshold signal to genome-wide significance. For example, the association of rs2814993 with leg mass, which has a suggestive $P = 1.1 \times 10^{-6}$ with standard GWAS, becomes GWS with SynSurr at $P = 2.3 \times 10^{-20}$ (in fact, this SNP is significant for all DEXA traits via SynSurr). rs2814993 is an intronic variant of the *ILRUN* gene, and was previously associated with height in a meta-analysis[42] of European populations and an Australian twin study[43]. As another example, rs17782313 is associated with all DEXA traits by SynSurr, foremost with total mass $P = 1.8 \times 10^{-26}$, but at best reaches a P of 1.6×10^{-5} with standard GWAS. rs17782313 is an intergenic variant near the *MC4R* gene, and has a well-characterized association with obesity[44-46].

Although we are not aware of an independent GWAS of the same traits, for external validation we overlapped SynSurr's findings with body composition associations from the GWAS Catalog[32]. On average, 70% of the DEXA associations identified by SynSurr were previously associated body composition in the GWAS Catalog (Fig. 6). As an internal validation, we performed a split-sample analysis, randomly allocating our GWAS data set 80:20 to independent discovery and validation cohorts. On average, 75.8% of GWS associations identified by SynSurr in the discovery cohort ($n = 277,998$) replicated in the validation cohort ($n = 69,500$)(Supplementary Table 22). Moreover, for all traits, the genetic correlation between the discovery and validation cohorts was high, 94.2% on average, with the 95% confidence interval always including 1.0 (Supplementary Fig. 21), underscoring the reproducibility of SynSurr's findings.

To investigate the biological function of the GWS variants, we performed gene set enrichment analysis using FUMA[47]. On average, 509 gene sets were enriched among the SynSurr results, while no significant enrichment was identified with standard GWAS. For all phenotypes, gene sets related to body fat distribution were among the most significant, and numerous enrichments related to anthropometric traits were identified (Supplementary Data 1).

Discussion

Here we introduced SynSurr, a robust and powerful approach to performing GWAS on a partially missing target outcome. Analyses of real and simulated data demonstrate that SynSurr estimates the proper genetic effect, controls the type I error, and provides power at least equaling, but typically exceeding, standard GWAS. Meanwhile, alternative approaches like proxy GWAS and multiple-imputation provide biased effect estimates, and either fail to control the type I error (proxy GWAS) or lack power (multiple-imputation).

For all analyses reported in the main text, the model-building and GWAS data sets were non-overlapping. For our UKBB analyses, the relatives of subjects in the GWAS data set formed a natural model-building data set. However, in data-limited settings, splitting the cohort into disjoint subsets will likely reduce power. Analyses in simulated (Extended Data Fig. 6) and real (Supplementary Tables 15-16) data suggest that SynSurr remains valid when the same subjects are reused for both model-building and GWAS. While dropping the requirement for disjoint model-building and GWAS data sets would simplify analyses and facilitate training more complex surrogate models, more theoretical work is needed to understand the circumstances under which this is justified.

Many additional analyses are reported in the Supplementary Materials. Supplementary Section 6 examines the trade-off between allocating subjects to the model-building versus GWAS data sets, given that the two will be kept disjoint. We find that allocating more of the labeled subjects to the GWAS data set improved power (Supplementary Fig. 5). Supplementary Section 7 examines the validity of imputing an input to the imputation model (i.e., the model that generates $\hat{Y}$). Unlike imputing Y, imputing a model input does not modify the relationship of interest (i.e., that between G and Y, as quantified by β_G). Thus, as demonstrated in Supplementary Fig. 6, imputing an input to $\hat{Y}$ neither biases the estimated genetic effects nor inflates the type I error. Supplementary Section 8 examines the bias introduced by selectively depleting the sample of subjects with extreme phenotypes, introducing missingness not at random (MNAR)[21]. Although all methods incur bias when the data are MNAR, SynSurr is no more sensitive to MNAR than standard or imputation-based GWAS (Supplementary Fig. 7, Supplementary Table 5).

Lastly, SynSurr is not without limitations. First, the method was derived assuming the joint distribution of the target phenotype and the synthetic surrogate is bivariate normal. To mitigate departures from this assumption, we suggest applying the rank-based inverse normal transformation[48] to both Y and $\hat{Y}$, and did so for all analyses. Our current work focuses on dropping the bivariate normality assumption. This can be achieved by replacing the score equations derived from maximum likelihood theory with a set of weighted estimating equations[49]. Moreover, we plan to extend SynSurr to generalized linear models with outcomes from the exponential family using a zero-mean augmentation approach[50,51]. Second, SynSurr requires that the target phenotype is missing at random (MAR; see Supplementary Methods). This assumption is prevalent in the GWAS literature (Extended Data Fig. 7), and is made by any study that performs complete-case analysis[21]. While MAR is expected to hold for the DEXA phenotypes, because the invitation to participate did not depend on a subject's body composition[39], it may fail in settings where the factors

affecting ascertainment are unknown[21]. A future direction to develop an implementation of SynSurr that is applicable when the phenotype is MNAR, perhaps by specifying models for the missing data mechanism, then performing various sensitivity analyses. Third, SynSurr does not currently accommodate related individuals. Future work will introduce a subject-specific random effect for modeling genetic relatedness[52]. Finally, SynSurr currently requires individual-level data. An important next step is to develop an implementation that works from summary statistics.

Methods

Statement on Ethics

Our use of data from the UK Biobank (UKBB; https://www.ukbiobank.ac.uk) was approved under application 64875, and our study complied with all conditions and access procedures set forth by the UK Biobank.

Statistics and Reproducibility

For the analyses in UKBB, all participants of self-identified ‘White British’ ancestry with available genetics and non-missing values for either the target outcome or the surrogate outcomes were included. Restriction to the largest single ancestry was solely for the purpose of avoiding spurious associations due to population structure. Among sets of kin, one subject was allocated at random to the GWAS data set, and the remaining subjects to the model-building data set. No statistical method was used to predetermine sample size. Study participants were not randomized, however, as discussed in the Mendelian randomization literature [53], the genotypes of unrelated individuals at loci not in linkage can be viewed as randomly assigned in the absence of confounding (e.g., due to population structure). The investigators did not ascertain any data from participants, and were not blinded to any data.

Standard GWAS

For the standard GWAS, the target phenotype Y is regressed on genotype G and covariates X among those subjects whose target phenotypes are observed:

$$Y = G\beta_G + X\beta_X + \epsilon\,. \tag{2}$$

The null hypothesis $H_0 : \beta_G = 0$ is evaluated using the standard two-sided Wald test[54]:

$$T_W = \frac{\hat{\beta}_G^2}{\mathbb{SE}^2(\hat{\beta}_G)}, \tag{3}$$

where $\hat{\beta}_G$ is the ordinary least squares (OLS) estimate of the genetic effect, and $\mathbb{SE}(\hat{\beta}_G)$ is the corresponding standard error.

SynSurr GWAS

As shown in Figure 1, we recommend allocating disjoint model-building and inference data sets. Within the inference data, suppose the target phenotype Y is observed for n_{obs} subjects and missing for n_{miss}, with $n = n_{\text{obs}} + n_{\text{miss}}$ denoting the total sample size. Denote by $\hat{Y}$ a synthetic surrogate for Y that is available for all n subjects. We recommend constructing $\hat{Y}$ by means of a nonlinear ML model, however Y could simply be another phenotype. Let G denote the genotype and X a vector of covariates, such as age, sex, and genetic principal components. To make use of both Y and $\hat{Y}$ in evaluating the association between Y and G, SynSurr utilizes the joint association model:

$$\begin{bmatrix} Y \\ \hat{Y} \end{bmatrix} \mid Z = \begin{bmatrix} Z & 0 \\ 0 & Z \end{bmatrix} \begin{bmatrix} \beta \\ \alpha \end{bmatrix} + \begin{bmatrix} \epsilon_T \\ \epsilon_S \end{bmatrix} \tag{4}$$

where $Z = (G, X)^T$, $\beta = (\beta_G, \beta_X)^T$, $\alpha = (\alpha_G, \alpha_X)^T$, and the residuals $(\epsilon_T, \epsilon_S)^T$ follow a bivariate normal distribution:

$$\begin{bmatrix} \epsilon_T \\ \epsilon_S \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{TT} & \Sigma_{TS} \\ \Sigma_{ST} & \Sigma_{SS} \end{bmatrix} \right).$$

Here, the subscripts of T and S denote the *target* phenotype and the *synthetic surrogate* respectively. In model (4), β_G is the parameter of interest, which quantifies the association between the genotype and the target phenotype. This is the same parameter estimated by standard GWAS of Y on G among the n_{obs} subjects with observed target phenotypes in (2). Proxy GWAS, which regresses $\hat{Y}$ rather than Y on G, estimates α_G instead of β_G. Needless to say, α_G can differ from β_G, and likely will when $\hat{Y}$ is generated from a misspecified model. SynSurr enables inference on β_G while making use of all n subjects. Moreover, SynSurr is computationally tractable at biobank scale, requiring only two ordinary least squares regressions:

1. First, among all n subjects, regress $\hat{Y}$ on Z to obtain an estimate of α.
2. Second, among the n_{obs} subjects with observed phenotypes, regress Y on $(\hat{Y}, Z)^T$ to obtain an estimate of the associated regression coefficient, denoted $(\delta, \gamma)^T$.

The validity of this two-step approach is demonstrated through a reparameterization of the log-likelihood function which allows the association parameter to be recovered as $\beta = \gamma + \delta\alpha$. Details of this equivalence, as well as derivation of the Wald test for SynSurr, are provided in the Supplementary Methods.

UK Biobank genotype and sample quality control

Our UKBB data release contains genotypes for 488,377 subjects and 784,256 directly genotyped variants. Prior to the analysis, we performed the following common quality control steps[55]:

1. Excluded individuals with > 10% missing genotypes.
2. Excluded SNPs with a genotyping rate < 90%.
3. Excluded SNPs with a Hardy-Weinberg Equilibrium $p < 10^{-5}$.
4. Excluded SNPs with MAF < 1%.
5. Included only who self-identified as 'White British' and have very similar genetic ancestry based on a principal components analysis of the genotypes (Data-Field 22006).
6. Selected one member of each set of subjects with a kinship coefficient greater than 0.0625 (the threshold for third-degree relatives) for the GWAS data set, and allocated the remaining subjects to the model-building data set.

The stages from our genotype QC are summarized in Supplementary Table 6. After QC, 435,468 directly genotyped genetic variants remained. Note that the model-building data set is allowed to contain related individuals, as there is not statistical requirement for independence.

Ablation studies

In total, 349,474 unrelated subjects were included in the height GWAS and 308,518 in the FEV1 GWAS. For each phenotype, sets of kin up to the third-degree were identified. One subject was randomly allocated to the GWAS data set and the remaining subjects to the model-building data set. Note that the subjects allocated to model-building could not otherwise participate in the GWAS as our bivariate association model (Equation 4) currently does not accommodate related individuals. The model-building data were used to construct a random forest for predicting the target phenotype on the basis of age, sex, and anthropomorphic measurements (see Supplementary Section 10.1).

Body composition GWAS

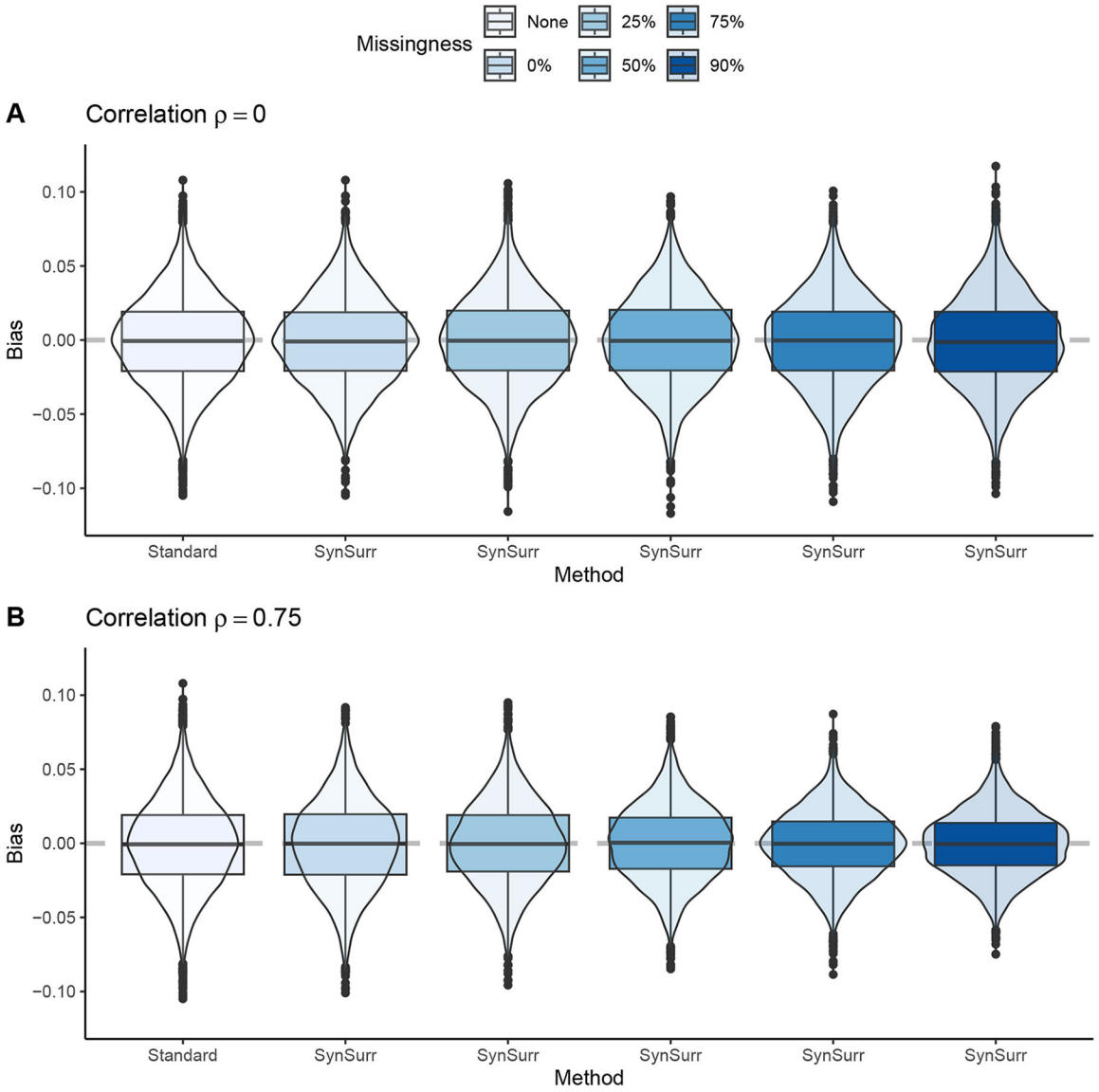
We performed SynSurr analyses of 6 incompletely measured DEXA phenotypes: android, arm, gynoid, leg, trunk, and total mass. In each case, the model-building data set included 4,584 subjects with observed target outcomes, while the GWAS data set included 347,498 subjects, 29,577 (8.5%) of which have observed target phenotypes. Within the model-building data set, a random forest was trained to predict the DEXA phenotype on the basis of age, sex, height, body weight, body mass index, and 5 measures of impedance: whole body, left/right arm, left/right leg. The fitted models were transferred to the GWAS data set, where a synthetic surrogate outcome was generated for all subjects.

GWAS catalog overlap analysis

Summary statistics for `body fat distribution, body fat percentage, fat body mass, lean body mass` were downloaded from the NHGRI-EBI GWAS catalog[32]. After concatenating and reducing to 1 record per unique combination of chromosome and base pair, this set contained 984 associated variants. Overlap of study variants with GWAS catalog variants was assessed using the `GenomicRanges` (v1.54.0)[56] package in `R`

(v4.3.2)[57]. A study variant was considered overlapped if fell within 250 kb of GWAS catalog variant for one of the aforementioned traits.

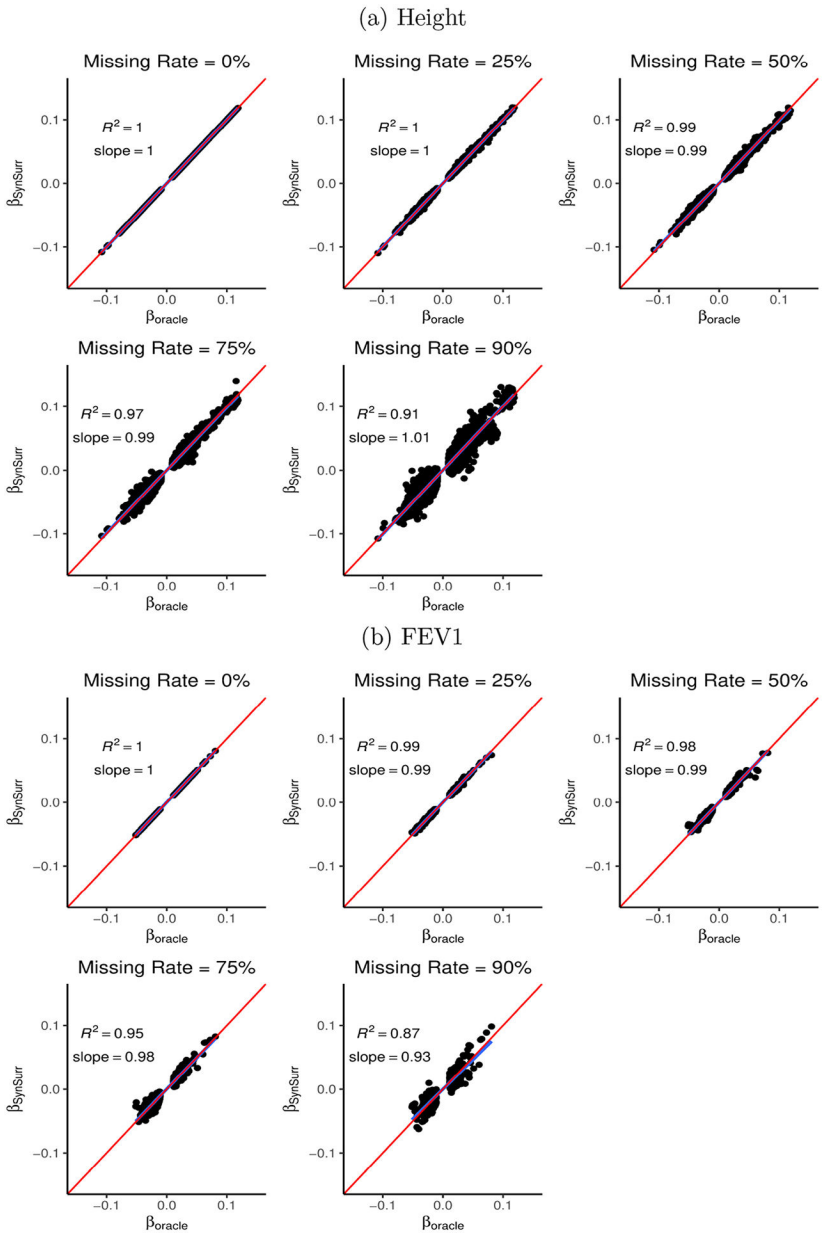
Extended Data



Extended Data Figure 1: Robustness and precision of SynSurr with an uninformative and informative synthetic surrogate.

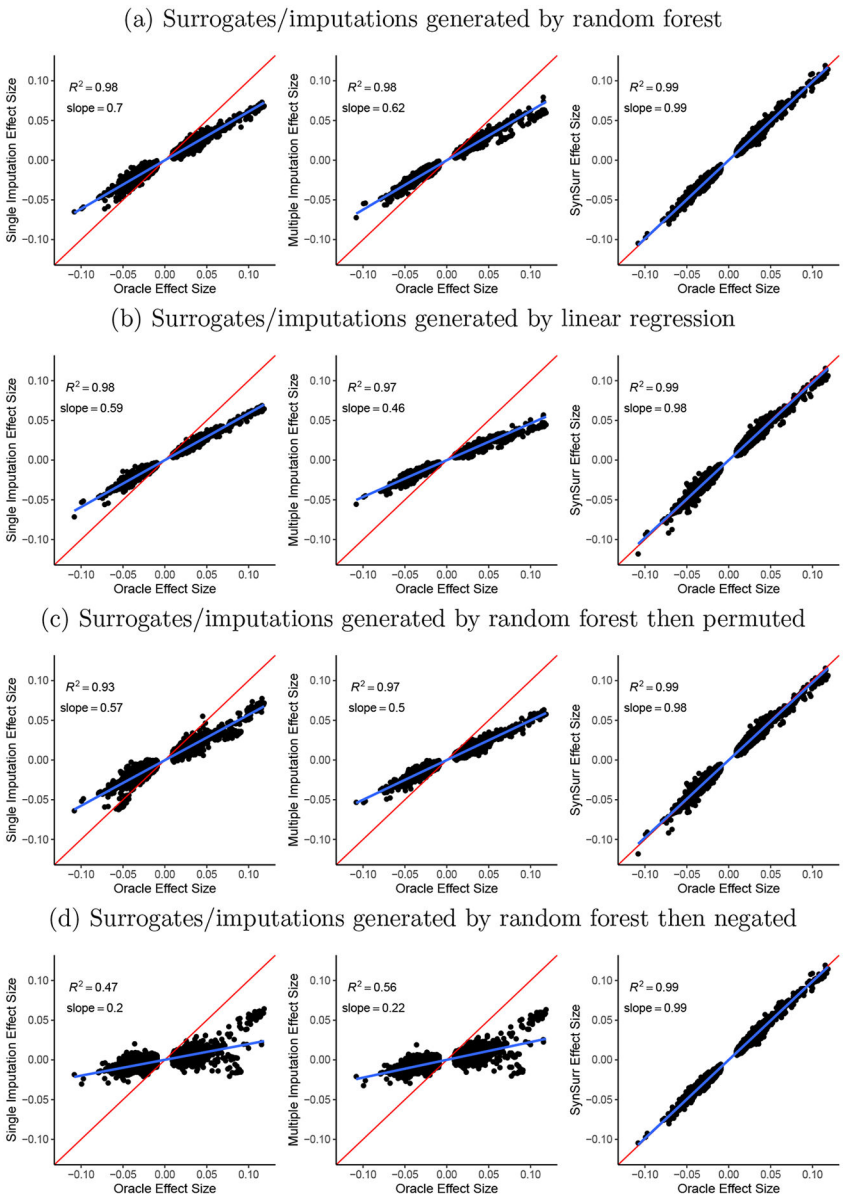
In all cases, the number of subjects with observed phenotypes was $n = 10^3$. The number of subjects with missing phenotypes was varied to achieve the indicated level of missingness. The standard estimator utilizes the observed values of Y only. In panel **A**, the synthetic surrogate has correlation $\rho = 0.00$ with the target phenotype, and is in fact independent of the target phenotype. Use of the SynSurr estimator with this uninformative surrogate results in no loss of efficiency relative to the standard analysis. In panel **B**, the synthetic surrogate has correlation $\rho = 0.75$ with the target phenotype. SynSurr becomes more efficient as the

number of subjects with missing target outcomes increases. The center of the box plot is the median, the upper and lower bounds of the box are the 75th and 25th percentiles, and the whiskers extend from the minimum to the maximum. The number of simulation replicates is 5×10^3.



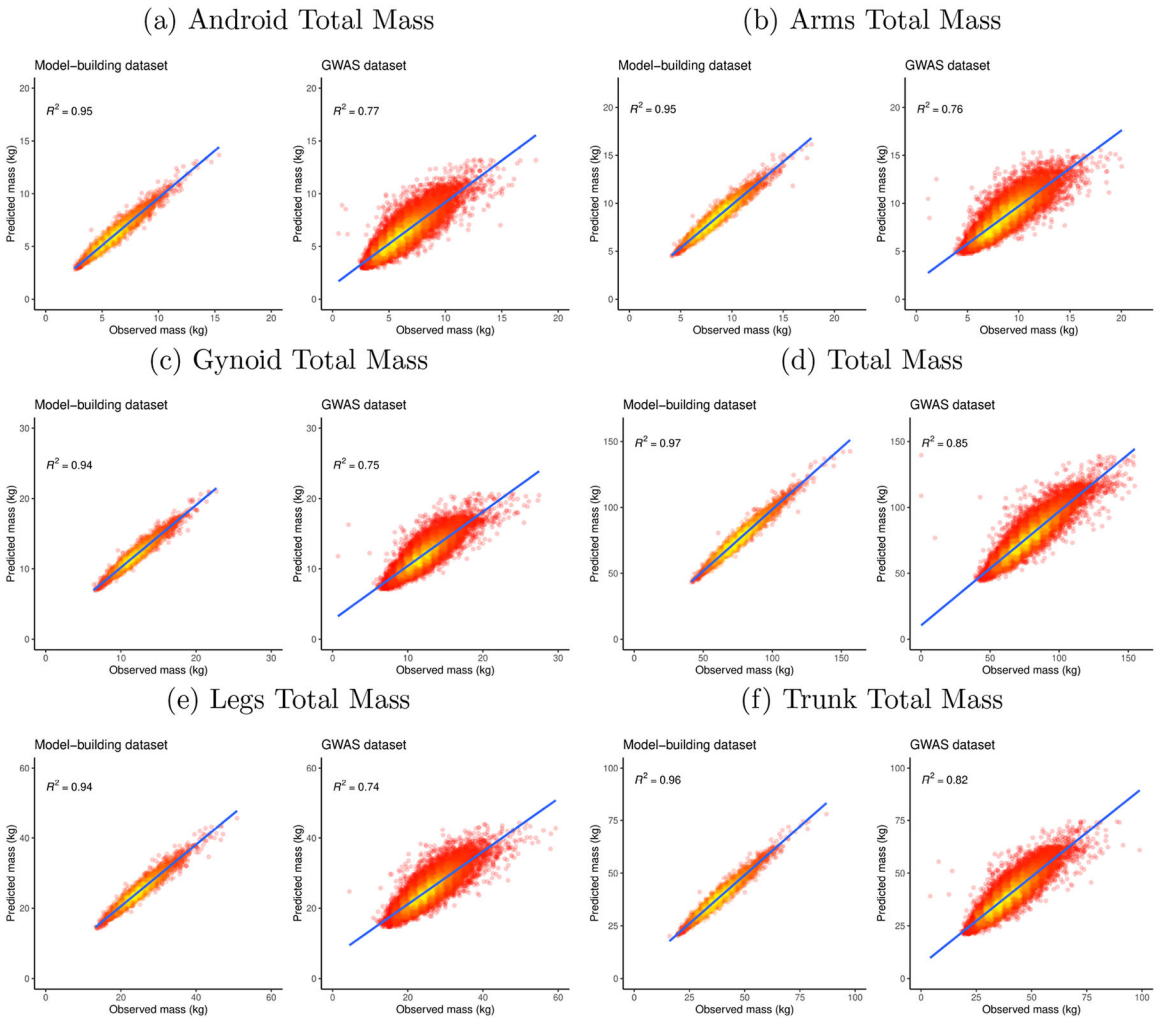
Extended Data Figure 2: Signal recovery of SynSurr relative to the oracle GWAS for height and FEV1.

A slope of 1.0 indicates that the estimated effect sizes are consistent with the oracle effect sizes. Note that although the slope deviates from 1.0 at 90% missingness, the slope approaches 1.0 as missingness declines. The following figure, which assesses signal recovery for standard GWAS, provides a point of comparison for the R^2 values.



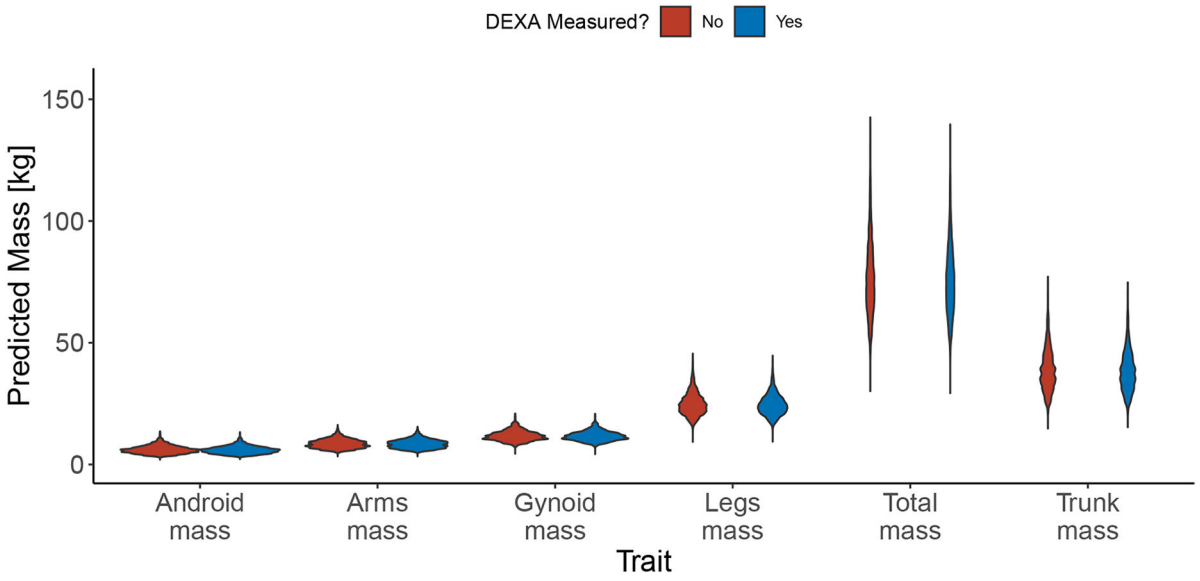
Extended Data Figure 3: Signal recovery of imputation-based approaches and SynSurr relative to the oracle GWAS for height with 50% missingness.

A slope of 1.0 indicates that the estimated effect sizes are consistent with the oracle effect sizes, whereas a slope deviating from 1.0 suggests the presence of bias.



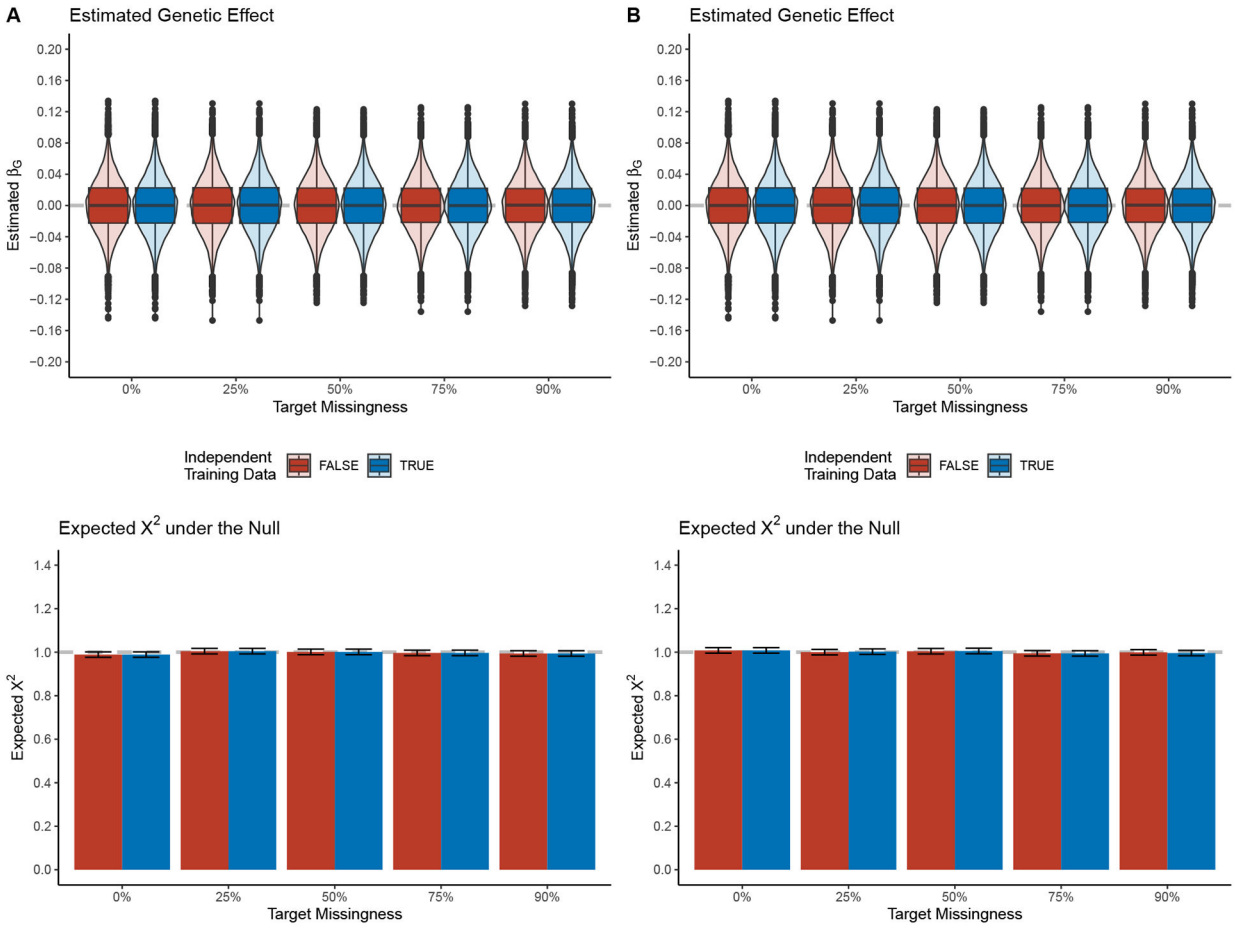
Extended Data Figure 4: Predicted vs. observed values of body composition phenotypes within the model-building and GWAS data sets.

A random forest was trained to predict each of the 6 body composition phenotypes, obtained via DEXA scan, using 4,584 subjects allocated to the model-building data set. The GWAS dataset consists of 29,577 unrelated subjects with body compositions measured via DEXA. Model inputs included age, sex, height, body weight, body mass index, and 5 impedance measures (whole body, left arm, right arm, left leg and right leg).



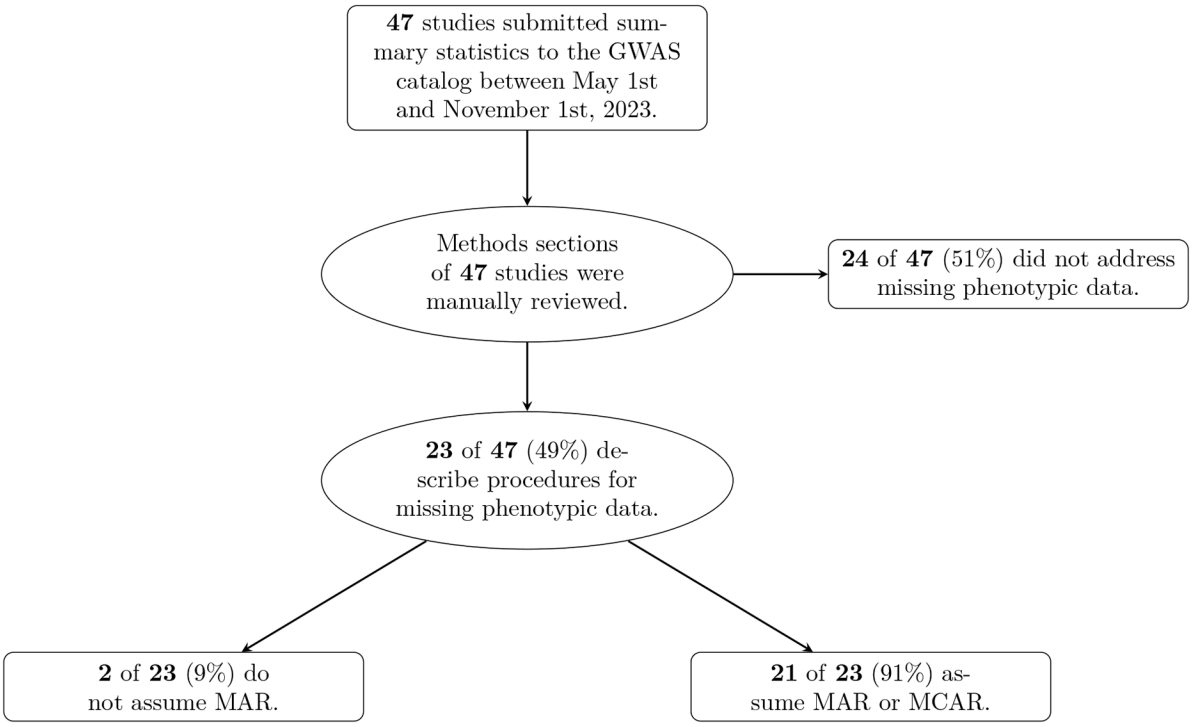
Extended Data Figure 5: Distribution of predicted body masses comparing subjects with and without DEXA measurements.

The violin plot shows the kernel density estimation of the distribution of the data, with the tips of the violin indicating the maximum and minimum observed values among subjects. Sample sizes: $n = 29,577$ independent subjects with DEXA measurements; $n = 317,921$ subjects without DEXA measurements.



Extended Data Figure 6: SynSurr remains unbiased and properly controls the type I error when the same data are utilized for model training and for GWAS.

The number of subjects with observed phenotypes was $n = 10^3$, while the number with missing phenotypes was varied to achieve the indicated level of missingness. The model that generated the synthetic surrogate was either trained in the GWAS data set or in an independent data set of size $n = 10^3$. Upper shows the distribution of effect sizes across 20×10^3. The true genetic effect size is $\beta_G = 0.1$. The center of the box plot is the median, the upper and lower bounds of the box are the 75th and 25th percentiles, and the whiskers extend from the 5th to the 95th percentile. Lower shows the average χ^2 statistic under $H_0 : \beta_G = 0$ across 50×10^3 simulation replicates, for which the expected value is 1.0. Error bars are 95% confidence intervals for the mean. Panel A (left) considers a "misspecified" ($k = 2$) model that can only capture quadratic dependence of Y on X, while Panel B (right) considers a "correctly specified" model ($k = 3$) that can capture the cubic dependence. As seen, the validity of SynSurr is not contingent on correct specification of the surrogate model.



Extended Data Figure 7: Survey of assumptions surrounding missing phenotypic data in GWAS.
The methods sections of all studies contributing summary statistics to the GWAS catalog between May 1st and November 1st, 2023, were manually reviewed. Among 47 studies, 24 did not address missing phenotypic data. Of the 23 remaining, 21 made an assumption of missing at random (MAR) or missing completely at random (MCAR).

Extended Data Table 1:
Type I error and power of SynSurr across various missing rates and synthetic surrogates.

In all cases, the number of subjects with observed phenotypes was $n = 10^3$, while the number with missing phenotypes was varied to achieve the indicated level of missingness. The synthetic surrogate has correlation $\rho = 0.00, 0.25, 0.50, 0.75$, 0.25, 0.50, 0.75 with the target phenotype. The power is reported for the setting with SNP heritability of 0.5% ($\beta_G = 0.07$). Type I error is controlled across all settings. The power is stable across values of ρ when there is no missing phenotype information, which is asymptotically equivalent to the standard analysis. The power of SynSurr increases with increasing missing rate and correlation. The number of simulation replicates is $R = 10^5$.

Missing Rate (%)	ρ	Type I Error	χ^2	Power	χ^2
0	0.00	0.05	1.01	0.65	6.58
0	0.25	0.05	1.00	0.65	6.59
0	0.50	0.05	0.99	0.65	6.56
0	0.75	0.05	1.00	0.65	6.59

Missing Rate (%)	ρ	Type I Error	χ^2	Power	χ^2
25	0.00	0.05	0.99	0.66	6.59
25	0.25	0.05	1.00	0.66	6.69
25	0.50	0.05	1.00	0.68	6.96
25	0.75	0.05	1.00	0.72	7.52
50	0.00	0.05	1.01	0.65	6.56
50	0.25	0.05	1.00	0.67	6.76
50	0.50	0.05	0.99	0.71	7.37
50	0.75	0.05	1.00	0.79	8.77
75	0.00	0.05	1.00	0.65	6.56
75	0.25	0.05	1.00	0.68	6.89
75	0.50	0.05	1.00	0.75	7.90
75	0.75	0.05	1.00	0.87	10.68
90	0.00	0.05	1.00	0.66	6.58
90	0.25	0.05	1.00	0.68	6.84
90	0.50	0.05	1.01	0.76	8.29
90	0.75	0.05	0.99	0.92	12.2

Extended Data Table 2:
Comparison of SynSurr with Proxy and MTAG GWAS for height.

Proxy GWAS analyzes the synthetic surrogate $\hat{Y}$ in place of the true target outcome Y. MTAG augments the results from standard GWAS, for a given level of missingness, with those from proxy GWAS. An association was considered a true positive if it was identified by the oracle GWAS (i.e. standard GWAS in the absence of missingness), and a false positive if it was not identified by the oracle GWAS. Below, “Total” is the total number of genome-wide significant associations, “True Positives” gives the number and percentage of oracle variants recovered, and “False Positives” gives the number and percentage of the total associations that were not detected by the oracle.

Method	Missing (%)	Oracle	Total	True Positives	False Positives
Proxy	0	7,177	1,654	1,363(18.99%)	291(17.59%)
MTAG	0	7,177	2,981	2,981(41.54%)	0(0%)
SynSurr	0	7,177	7,177	7,177(100%)	0(0%)
MTAG	25	7,177	2,384	2,384(33.22%)	0(0%)
SynSurr	25	7,177	5,416	5,305(73.92%)	111(2.05%)
MTAG	50	7,177	1,703	1,698(23.66%)	5(0.29%)
SynSurr	50	7,177	3,453	3,421(47.67%)	32(0.93%)
MTAG	75	7,177	1,075	1,057(14.73%)	18(1.67%)
SynSurr	75	7,177	1,250	1,243(17.32%)	7(0.56%)
MTAG	90	7,177	832	788(10.98%)	44(5.29%)

Method	Missing (%)	Oracle	Total	True Positives	False Positives
SynSurr	90	7,177	329	329(4.58%)	0(0%)

Extended Data Table 3:
Comparison of genome-wide significant SNPs discovered by SynSurr with Standard GWAS for the UKBB DEXA phenotypes.

DEXA Phenotype	SynSurr Only	SynSurr and Standard	Standard Only
Android	65	8	0
Arms	80	9	1
Gynoid	252	8	0
Legs	270	8	0
Total	268	8	0
Trunk	142	8	0

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Institutes of Health grants R35-CA197449 and F31-HL140822 (to ZRM); and R35-CA197449, U19-CA203654, R01-HL163560, U01-HG012064, and U01-HG009088 (to XL); and the Natural Sciences and Engineering Research Council of Canada grant RGPIN-2021-03734 and a Connaught New Researcher Award (to JeG). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability

This work used genotypes and phenotypes from the UK Biobank (https://www.ukbiobank.ac.uk). Summary statistics from the DEXA trait analysis will be deposited with the GWAS catalog, and are available upon reasonable request in the interim.

Code availability

`SurrogateRegression` (v0.6.0.1) is available as an R[57] package on the Comprehensive R Archive Network: https://CRAN.R-project.org/package=SurrogateRegression[58]. Replication code for the analyses presented in this paper is available on GitHub at: https://github.com/jianhuig/SyntheticSurrogateAnalysis[59].

References

1. Kurki M, Karjalainen J, Palta P, et al. FinnGen provides genetic insights from a well-phenotyped isolated population. Nature 613, 508–518 (2023). [PubMed: 36653562]
2. Gaziano JM et al. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. Journal of clinical epidemiology 70, 214–223 (2016). [PubMed: 26441289]

3. Bycroft C. et al. The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203–209 (2018). [PubMed: 30305743]
4. Beesley LJ et al. The emerging landscape of health research based on biobanks linked to electronic health records: Existing resources, statistical challenges, and potential opportunities. Statistics in medicine 39, 773–800 (2020). [PubMed: 31859414]
5. Tan VY & Timpson NJ The UK Biobank: A Shining Example of Genome-Wide Association Study Science with the Power to Detect the Murky Complications of Real-World Epidemiology. Annual Review of Genomics and Human Genetics 23 (2022).
6. Wei W-Q & Denny JC Extracting research-quality phenotypes from electronic health records to support precision medicine. Genome medicine 7, 1–14 (2015). [PubMed: 25606059]
7. Banda JM, Seneviratne M, Hernandez-Boussard T & Shah NH Advances in electronic phenotyping: from rule-based definitions to machine learning models. Annual review of biomedical data science 1, 53 (2018).
8. Allen N, Sudlow C, Peakman T, Collins R & Biobank U UK biobank data: come and get it. Sci Transl Med 6, 224ed4 (2014).
9. Littlejohns TJ et al. The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. Nature communications 11, 1–12 (2020).
10. Elliott L. et al. Genome-wide association studies of brain imaging phenotypes in UK Biobank. Nature 562, 210–216 (2018). [PubMed: 30305740]
11. Pirruccello J. et al. Analysis of cardiac magnetic resonance imaging in 36,000 individuals yields genetic insights into dilated cardiomyopathy. Nat Commun 11, 2254 (2020). [PubMed: 32382064]
12. Alipanahi B, Hormozdiari F, Behsaz B, et al. Large-scale machine-learning-based phenotyping significantly improves genomic discovery for optic nerve head morphology. American Journal of Human Genetics 108, 1217–1230 (2021). [PubMed: 34077760]
13. Li X & Zhao H Automated feature extraction from population wearable device data identified novel loci associated with sleep and circadian rhythms. PLoS Genet 16, e1009089 (2020). [PubMed: 33075057]
14. Hormozdiari F. et al. Imputing phenotypes for genome-wide association studies. The American Journal of Human Genetics 99, 89–103 (2016). [PubMed: 27292110]
15. Dahl A, Iotchkova V, Baud A, et al. A multiple-phenotype imputation method for genetic studies. Nature Genetics 48, 466–472 (2016). [PubMed: 26901065]
16. Zhang Y. et al. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). Nature protocols 14, 3426–3444 (2019). [PubMed: 31748751]
17. Liao KP et al. High-throughput multimodal automated phenotyping (MAP) with application to PheWAS. Journal of the American Medical Informatics Association 26, 1255–1262 (2019). [PubMed: 31613361]
18. Cosentino J, Behsaz B, Alipanahi B, et al. Inference of chronic obstructive pulmonary disease with deep learning on raw spirograms identifies new genetic loci and improves risk models. Nature Genetics 55, 787–795 (2023). [PubMed: 37069358]
19. An U, Pazokitoroudi A, Alvarez M, et al. Deep learning-based phenotype imputation on population-scale biobank data increases genetic discoveries. Nature Genetics 55, 2269–2276 (2023). [PubMed: 37985819]
20. Dahl A, Thompson M, An U, et al. Phenotype integration improves power and preserves specificity in biobank-based genetic studies of major depressive disorder. Nature Genetics 55, 2082–2093 (2023). [PubMed: 37985818]
21. Little RJ & Rubin DB Statistical Analysis with Missing Data 2nd (John Wiley & Sons, 2002).
22. Wang S, McCormick TH & Leek JT Methods for correcting inference based on outcomes predicted by machine learning. Proceedings of the National Academy of Sciences 117, 30266–30275 (2020).
23. Hubbard RA, Tong J, Duan R & Chen Y Reducing bias due to outcome misclassification for epidemiologic studies using EHR-derived probabilistic phenotypes. Epidemiology 31, 542–550 (2020). [PubMed: 32282406]

24. Hong C, Liao KP & Cai T Semi-supervised validation of multiple surrogate outcomes with application to electronic medical records phenotyping. Biometrics 75, 78–89 (2019). [PubMed: 30267536]
25. Rubin D. Multiple Imputation for Nonresponse in Surveys (John Wiley & Sons, 1987).
26. Rubin DB Multiple imputation after 18+ years. Journal of the American statistical Association 91, 473–489 (1996).
27. Van Buuren S. Flexible Imputation of Missing Data 2nd (Chapman and Hall/CRC, 2018).
28. Bartlett JW & Hughes RA Bootstrap inference for multiple imputation under uncongeniality and misspecification. Statistical methods in medical research 29, 3533–3546 (2020). [PubMed: 32605503]
29. Austin PC, White IR, Lee DS & van Buuren S Missing data in clinical research: a tutorial on multiple imputation. Canadian Journal of Cardiology 37, 1322–1331 (2021). [PubMed: 33276049]
30. Murray JS Multiple imputation: a review of practical and theoretical findings. Statistical Science 33, 142–159 (2018).
31. McCaw ZR, Gaynor SM, Sun R & Lin X Leveraging a surrogate outcome to improve inference on a partially missing target outcome. *Biometrics* **Online ahead of print** (2022).
32. Buniello A, MacArthur J, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 47, D1005–D1012 (2019). [PubMed: 30445434]
33. Breiman L. Random forests. Machine learning 45, 5–32 (2001).
34. Chen T & Guestrin C XGBoost: A Scalable Tree Boosting System. *CoRR* **abs/1603.02754.** arXiv: 1603.02754. http://arxiv.org/abs/1603.02754 (2016).
35. Casella B & Berger R Statistical Inference. 2nd ed. (Duxbury/Thomson Learning, Pacific Grove, CA, 2002).
36. Rubin DB Inference and missing data. Biometrika 63, 581–592 (1976).
37. Allen NE, Sudlow C, Peakman T, Collins R & biobank U UK biobank data: come and get it 2014.
38. Biobank U. UK Biobank Body Composition Measurement https://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=1421. 2011.
39. Littlejohns T, Holliday J, Gibson L, et al. The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. Nat Commun 11, 2624 (2020). [PubMed: 32457287]
40. Biobank U. UK Biobank Imaging Modality DXA https://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=502. 2015.
41. Turley P, Walters R, Maghzian O, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. Nature Genetics 50, 229–237 (2018). [PubMed: 29292387]
42. Weedon M, Lango H, Lindgren C, et al. Genome-wide association analysis identifies 20 loci that influence adult height. Nature Genetics 40, 575–583 (2008). [PubMed: 18391952]
43. Liu J, Medland S, Wright M, et al. Genome-wide association study of height and body mass index in Australian twin families. Twin Res Hum Genet 13, 179–193 (2010). [PubMed: 20397748]
44. Meyre D, Delplanque J, Chevre J, et al. Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. Nature Genetics 41, 157–159 (2009). [PubMed: 19151714]
45. Willer C, Speliotes E, Loos R, et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nature Genetics 41, 25–34 (2009). [PubMed: 19079261]
46. Loos R & Yeo G The genetics of obesity: from discovery to biology. Nat Rev Genet 23, 120–133 (2022). [PubMed: 34556834]
47. Watanabe K, Taskesen E, van Bochoven A & Posthuma D Functional mapping and annotation of genetic associations with FUMA. Nat Commun 8, 1826 (2017). [PubMed: 29184056]
48. McCaw Z, Lane J, Saxena R, Redline S & Lin X Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. Biometrics 76, 1262–1272 (2020). [PubMed: 31883270]

49. Robins J & Rotnitzky A Semiparametric Efficiency in Multivariate Regression Models with Missing Data. Journal of the American Statistical Association 90, 122–129 (1995).

50. Wang X & Wang Q Semiparametric linear transformation model with differential measurement error and validation sampling. Journal of Multivariate Analysis 141, 67–80 (2015).

51. Tong J. et al. An augmented estimation procedure for EHR-based association studies accounting for differential misclassification. Journal of the American Medical Informatics Association 27, 244–253 (2020). [PubMed: 31617899]

52. Po-Ru L, Tucker G, Bulik-Sullivan B, Vilhjalmsson B, Finucane H, et al. Efficient Bayesian mixed model analysis increases association power in large cohorts. Nature Genetics 47, 284–290 (2015). [PubMed: 25642633]

54. Seber G. The Linear Model and Hypothesis. A General Unifying Theory 1st ed. (Springer Cham, 2015).

55. Purcell S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 81, 559–575 (2007). [PubMed: 17701901]

56. Lawrence M, Huber W, Pages H, et al. Software for Computing and Annotating Genomic Ranges. PLoS Comput Biol 9, e1003118 (2013). [PubMed: 23950696]

57. R Core Team. R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing (Vienna, Austria, 2022). https://www.R-project.org/.

60. Cox D. A note on data-splitting for the evaluation of significance levels. Biometrika 62, 441–444 (1975).

Methods-only References

53. Lawlor D, Harbord R, Sterne J, Timpson N & Smith G Mendelian randomization: Using genes as instruments for makingcausal inferences in epidemiology. Statistics in Medicine 27, 1133–1163 (2008). [PubMed: 17886233]

58. McCaw Z. Surrogate Regression 10.5281/zenodo.10897842.

59. Gao J, Gronsbell J & McCaw Z Synthetic Surrogate Analysis 10.5281/zenodo.10901237.

61. McCaw ZR SurrogateRegression: Surrogate Outcome Regression Analysis Comprehensive R Archive Network (2020). https://CRAN.R-project.org/package=SurrogateRegression.

62. Meng X-L & Rubin DB Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika 80, 267–278 (1993).

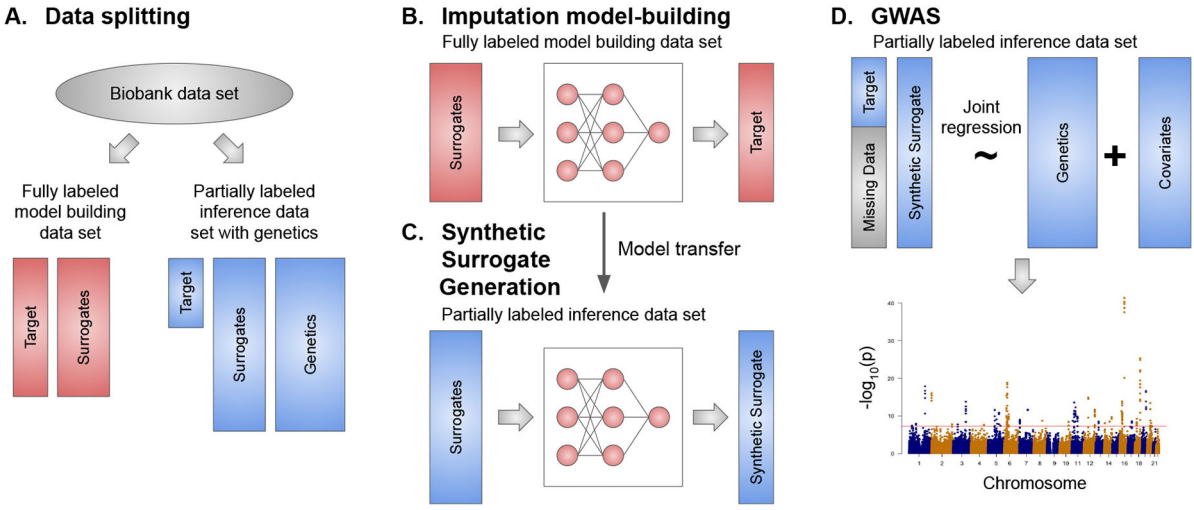


Figure 1: Graphical overview of a SynSurr GWAS.

A. The data set is first split into a fully labeled model-building data set, including the target phenotype and surrogates, and a partially labeled inference data set, which also includes genetics. **B.** Within the model-building data set, an imputation model is trained to predicted the target phenotype on the basis of surrogates. **C.** The imputation model is transferred to the partially labeled inference data set and applied to predict the target outcome for all subjects. The predicted value of the target outcome is referred to as the "synthetic surrogate". Importantly, the synthetic surrogate is maintained as a separate and distinct outcome from the partially missing target phenotype. **D.** Finally, within the inference data set, the partially missing target phenotype and the fully observed synthetic surrogate are jointly regressed on genotype and covariates to identify genetic variants associated with the target outcome.

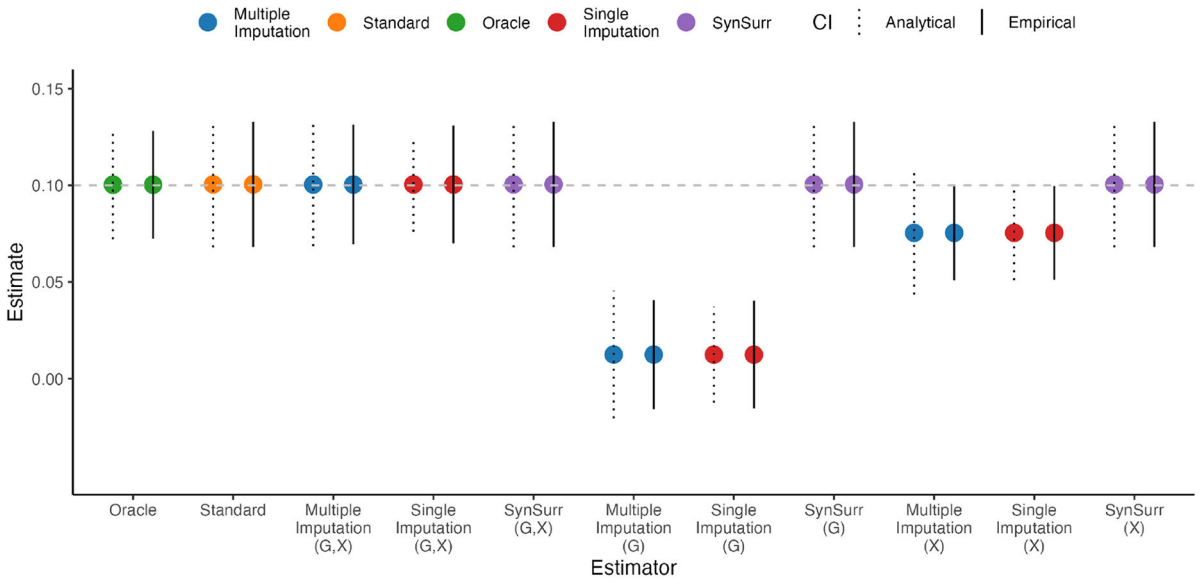


Figure 2: Unlike imputation-based estimators, SynSurr is robust to misspecification of the imputation model.

The true value for the parameter of interest is $\beta_G = 0.1$, corresponding to a variant with $h^2 = 1\,\%$. For each estimator, the sample size is $n = 10^4$, the mean value across 10^3 simulations is shown by the point, and two 95% confidence intervals (CIs) are presented: the dotted CI is based on the analytical standard error (SE) while the solid CI is based on the empirical SE. The oracle estimator has access to the complete version of Y, before 25% of values were set to missing. The standard estimator has access to the observed values of Y only. The imputation-based estimators impute the missing values of Y using an imputation-model fit on an independent data set. The set of covariates used to fit the imputation model are shown as a tuple: the imputation model based on G and X is correctly specified, whereas that based on G alone or X alone is misspecified. The SynSurr estimator jointly analyzes the partially missing Y with the synthetic surrogate $\hat{Y}$, where $\hat{Y}$ is generated for all subjects from the imputation model. The key observation is that SynSurr does not require a correctly specified generative model to yield unbiased estimation and valid inference.

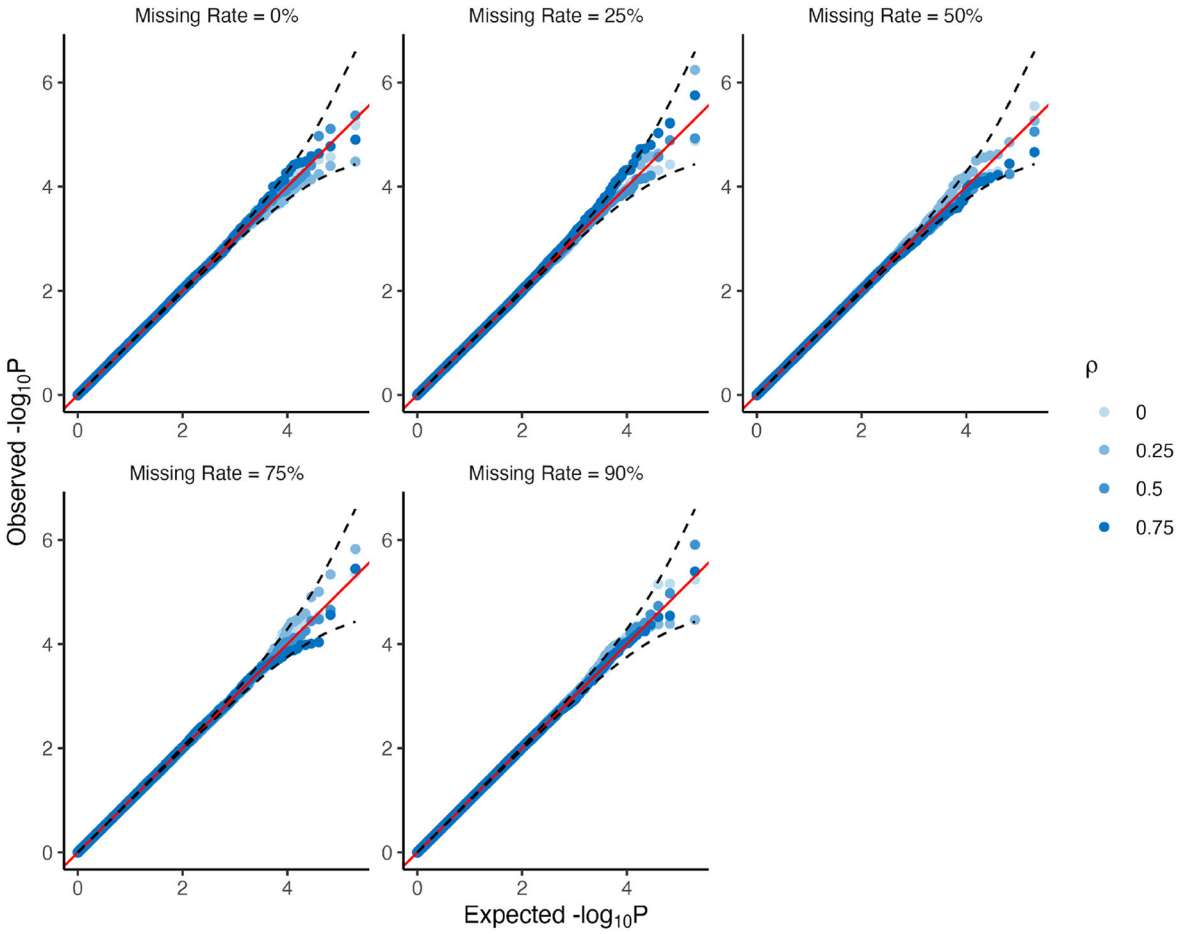


Figure 3: SynSurr controls type I error across missingness rates and target-surrogate correlations.

Type I error is the probability of incorrectly rejecting the null hypothesis $H_0 : \beta_G = 0$. In all cases, the number of subjects with observed phenotypes was $n = 10^3$. The number of subjects with missing phenotypes was varied to achieve the indicated level of missingness. The synthetic surrogate has correlation $\rho \in \{0.00, 0.25, 0.50, 0.75\}$ with the target phenotype. The number of simulation replicates is 10^6. P-values are two-sided and were calculated by SynSurr. Error bands (dashed black lines) represent 95% confidence intervals around the expected $-\log_{10}$ (p-values) under the null hypothesis. Adherence to the diagonal (red line) indicates that the p-values are uniformly distributed under the null.

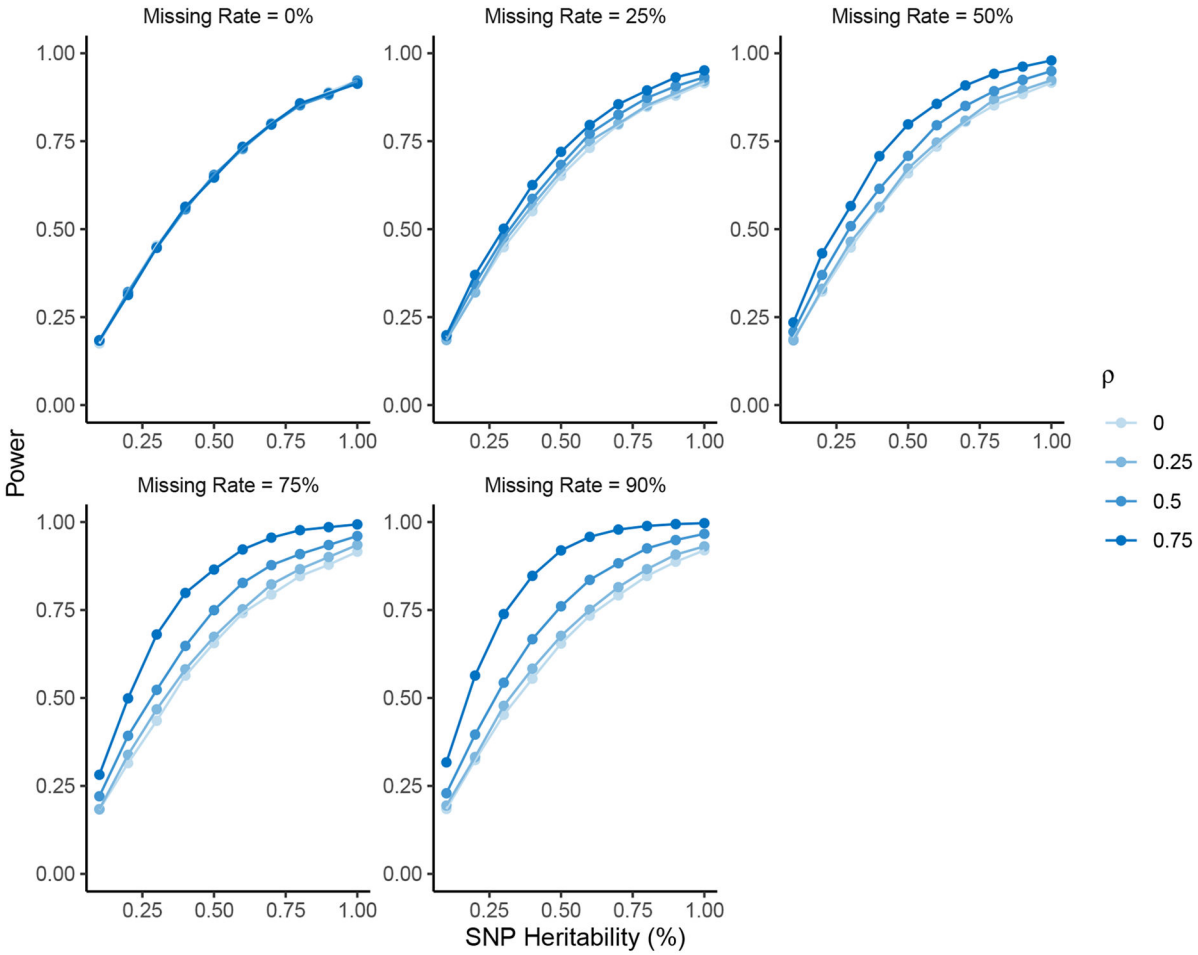


Figure 4: Power of SynSurr across various missing rates, target-surrogate correlations, and SNP heritabilities.

Power is the probability of correctly rejecting the null hypothesis $H_0 : \beta_G = 0$. In all cases, the number of subjects with observed phenotypes was $n = 10^3$. The number of subjects with missing phenotypes was varied to achieve the indicated level of missingness. In each panel, the synthetic surrogate has correlation $\rho \in \{0.00, 0.25, 0.50, 0.75\}$ with the target phenotype and the SNP heritability was varied from 0.1% to 1%. When there is no missingness, SynSurr is equivalent to the standard analysis and shows no variation across values of ρ. The power of SynSurr increases with increasing missingness and target-surrogate correlation. The number of simulation replicates is 10^4.

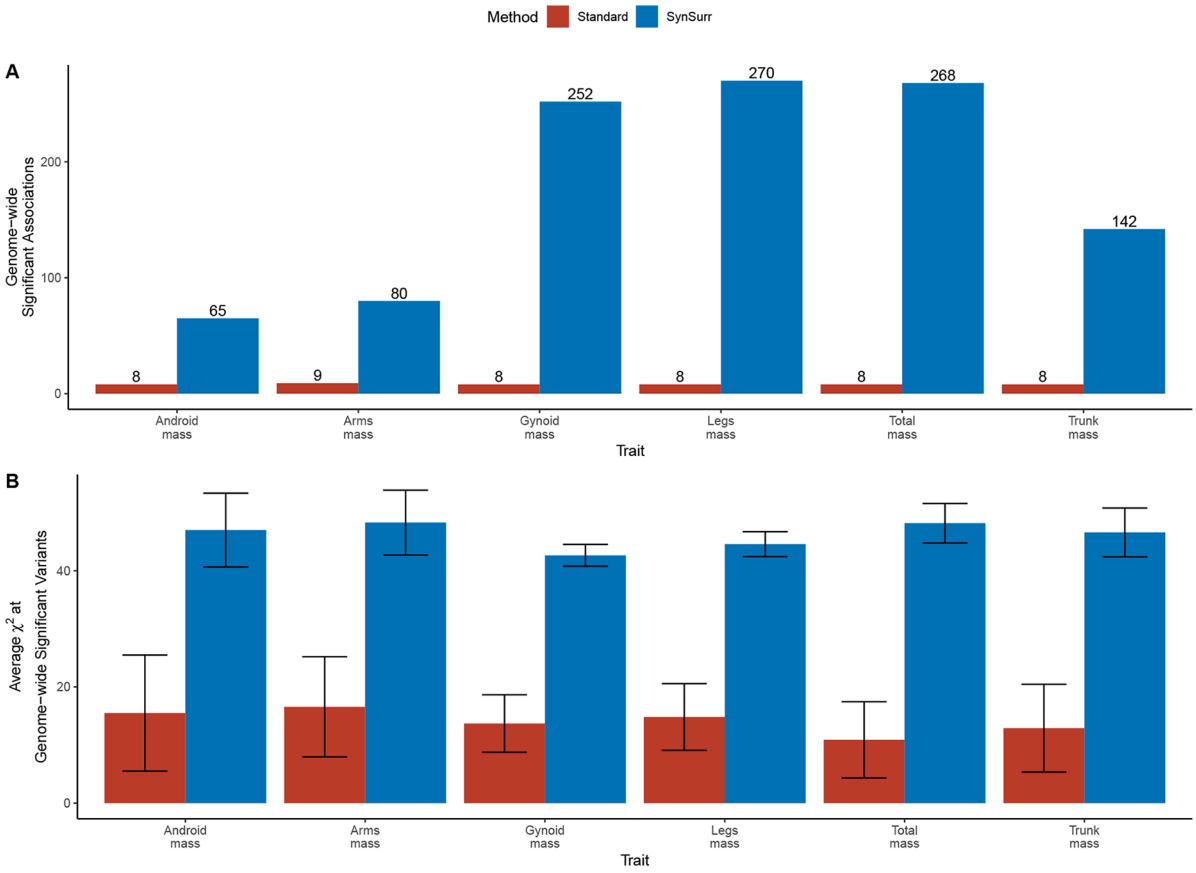


Figure 5: Comparing SynSurr and standard GWAS with respect to the number and significance of genome-wide significant associations for body composition traits.

A. Number of genome-wide significant (GWS) associations ($p < 5 \times 10^{-8}$) with DEXA body composition traits for standard and synthetic surrogate (SynSurr) GWAS. P-values are two-sided and are calculated by linear regression (Standard) or SynSurr. **B.** Average χ^2 statistic at the union of variants that reached genome-wide significance under either method. A greater expected χ^2 statistic directly corresponds to greater power to detect an association. Error bars are 95% confidence intervals for the mean. The number of independent GWS variants averaged across is shown in A.

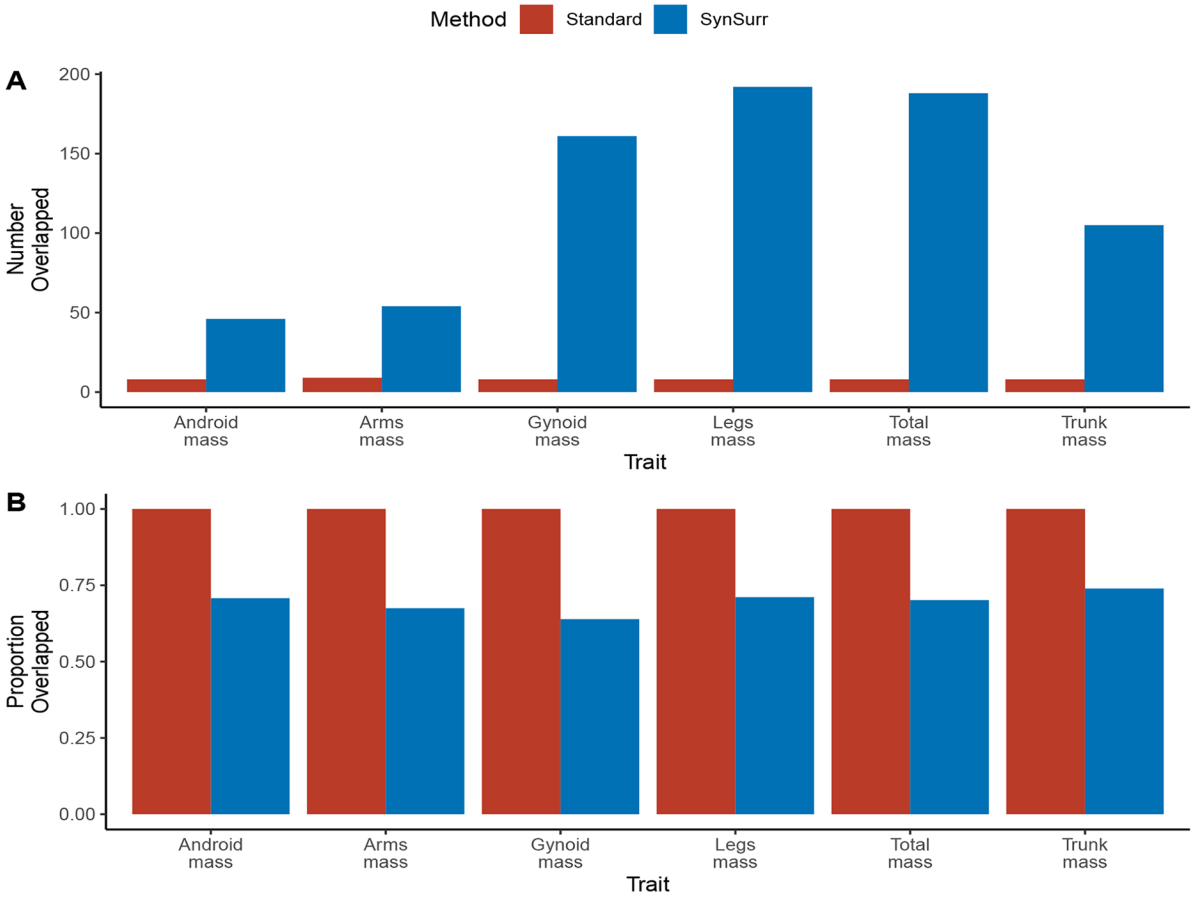


Figure 6: External validation via overlap of genome-wide significant variants for body composition with associations from the GWAS catalog.

Variants from the GWAS catalog associated with body fat distribution, body fat percentage, fat body mass, and lean body mass were compiled. A study variant was considered overlapped if it fell within 250 kb of a GWAS catalog variant. Panels A and B show the counts and proportions of overlapped variants, respectively. Note that, with 1 exception, all variants identified by standard GWAS were also identified by SynSurr (Supplementary Table 21). The perfect overlap of the standard GWAS variants with known body composition associations in panel B is a direct consequence of the standard GWAS detecting very few genome-wide significant variants (8.3 on average), and indicates that all of these variants were previously known.

Table 1:

Number of genome-wide significant SNPs recovered by standard and SynSurr GWAS across increasing ablation of the target phenotype.

The oracle method establishes the number of genome-wide significant (GWS) variants ($p < 5 \times 10^{-8}$) that would be identified in the absence of missingness. P-values are two-sided and calculated by linear regression (Oracle, Standard) or SynSurr. Missingness was introduced by ablating 25%, 50%, 75%, and 90% of the target phenotypes. Standard and SynSurr GWAS were performed on each of the ablated data sets. Standard GWAS refers to performing GWAS using only the observed values of the target outcome. The number and percentage of full-sample (oracle) GWS variants recovered by each method are reported. The false negative rate is 100% minus the recovery rate shown. Also see Supplementary Tables 7-9.

	Height				FEV1			
Missing Rate (%)	n_{obs}	**Oracle**	**Standard**	**SynSurr**	n_{obs}	**Oracle**	**Standard**	**SynSurr**
0	349,474	7,177	7,177(100%)	7,177(100%)	308,518	974	974(100%)	974(100%)
25	262,105	7,177	4,896(68.22%)	5,305(73.92%)	231,388	974	546(56.06%)	599(61.50%)
50	174,737	7,177	2,742(38.21%)	3,421(47.67%)	154,259	974	278(28.54%)	326(33.47%)
75	87,368	7,177	834(11.62%)	1,243(17.32%)	77,129	974	60(6.16%)	81(8.32%)
90	34,947	7,177	192(2.68%)	329(4.58%)	30,852	974	0(0%)	0(0%)