Impact of Measurement Noise on Genetic Association Studies of Cardiac Function*

Milos Vukadinovic[1,2,4], Gauri Renjith[1], Victoria Yuan[1,2], Alan Kwan[1,4], Susan C. Cheng[1], Debiao Li[4], Shoa L. Clarke[5,†], David Ouyang[1,6,†]

1. *Department of Cardiology, Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, CA*
2. *Department of Bioengineering, University of California Los Angeles, Los Angeles, CA*
3. *Department of Medicine, University of California Los Angeles, Los Angeles, CA*
4. *Biomedical Imaging Research Institute, Cedars-Sinai Medical Center, Los Angeles, CA*
5. *Department of Medicine, Division of Cardiovascular Medicine, Stanford University, Stanford, CA*
6. *Division of Artificial Intelligence in Medicine, Cedars-Sinai Medical Center, Los Angeles, CA*

†. Co-senior author

Email: shoa@stanford.edu, David.ouyang@cshs.org

1 Abstract

Recent research has effectively used quantitative traits from imaging to boost the capabilities of genome-wide association studies (GWAS), providing further understanding of disease biology and various traits. However, it's important to note that phenotyping inherently carries measurement error and noise that could influence subsequent genetic analyses. The study focused on left ventricular ejection fraction (LVEF), a vital yet potentially inaccurate quantitative measurement, to investigate how imprecision in phenotype measurement affects genetic studies. Several methods of acquiring LVEF, along with simulating measurement noise, were assessed for their effects on ensuing genetic analyses. The results showed that by introducing just 7.9% of measurement noise, all genetic associations in an LVEF GWAS with almost forty thousand individuals could be eliminated. Moreover, a 1% increase in mean absolute error (MAE) in LVEF had an effect equivalent to a 10% reduction in the sample size of the cohort on the power of GWAS. Therefore, enhancing the accuracy of phenotyping is crucial to maximize the effectiveness of genome-wide association studies.

Keywords: Precision phenotyping; Genome-Wide association study; Left ventricular ejection fraction; Cardiac magnetic resonance imaging; UK Biobank

* This work is partially supported by the National Institutes of Health NIH K99 HL157421.

2 Introduction

Cardiovascular disease is the leading cause of death in the world, and significant work has been undertaken to understand the mechanisms of disease and develop preventive measures. By studying the human genome, insights have been obtained to understand pathways and mechanisms of function and disease risk, and in recent studies, researchers have moved beyond binary labels of disease diagnosis to quantitative phenotypes to obtain greater power in assessing the relationship between genotype and phenotype[1–4]. From quantitative laboratory biomarkers elucidating the relationship between hypercholesterolemia and coronary artery disease[5] to imaging characteristics in population cohorts[4] revealing the genetic determinants of cardiovascular development [6,7], quantitative assessments of health provide additional signal compared to conventional binary labels of disease.

Despite its relative frequency, critical public health importance, and often penetrant inheritance, heart failure has relatively few known genetic risk factors. Early classic genetic studies were not able to identify many genetic associations with measurements determined by echocardiography[8]. Recent studies with larger cohorts and measurements from cardiac MRI have been able to find additional loci of relevance and reaffirm previously suspected variants[2], suggesting both larger sample sizes, as well as improvements in phenotyping precision, can improve our understanding of the human disease.

While quantitative traits often have more power than binary labels of disease, the issue of measurement error in quantitative traits is a known problem[9]. For example, left ventricular ejection fraction (LVEF) as measured by echocardiography can have measurement variation up to 7 - 10%[10,11], impacting downstream analyses. We use LVEF, the most prevalent metric of cardiac function, as an example of an important but noisy measurement to explore the impact of measurement variability on downstream genetic association studies. We compare various methods to obtain the same phenotypic measurement as well as introduce simulated noise in the phenotype measurement to evaluate the relative impact of measurement noise and sample size on downstream genetic studies.

Table 1. Cohort baseline characteristics

Characteristic	Mean or n
N	39624
Age at MRI	54.9 ± 7.47
Male	18933 (47.8%)
Self-identified White British	33726 (85.1%)
Body mass index (kg/m^2)	26.5 ± 4.19
Hypertension	2487 (6.3%)
Pulse rate	67.9 ± 10.9
LV ejection fraction (%)	55.4 (6.78)
LV end diastolic volume (mL)	141
LV end systolic volume (mL)	64.1

3 Methods

3.1 Cohort

The UK Biobank is a population-based cohort that links genetic and phenotypic data for approximately 500,000 adult participants from the United Kingdom [12,13]. We focused on 39,624 participants who had InlineVF measured LVEF [14], cardiac MRI, and genetic data available. Before running Genome-Wide Association Studies this cohort was passed through additional quality check filters **(Figure A1)**.

3.2 Multiple Approaches to Measure LVEF

Multiple methods of calculating LVEF from the same underlying imaging data were used to assess the impact of phenotyping precision on downstream analyses. First, the UKB provides automated LVEF measurements derived from MRI using Inline VF software[15], however, this is presented without manual quality control. To compare alternative automated approaches, we also derived LVEF from MRIs using the deep learning segmentation approach suggested by Bai et al [6]. From the short-axis view videos, segmentation was performed, we calculated the LV volume for each frame with Simpson's method and used the following LVEF formula:

$$\frac{ED\ Volume - ES\ Volume}{ED\ Volume} \times 100 \quad (1)$$

To simulate reader variability, additional experiments were performed introducing Gaussian noise with a mean of 0 and a standard deviation (sd) ranging from [1,10]. We generated multiple phenotypic measurements from the same underlying imaging data, gradually incrementing Gaussian noise, and performed GWAS on each to investigate how measurement error/imprecision affects genetic associations.

Additionally, we further compared results with two final approaches to assess LVEF. When visually assessing LVEF, clinicians often round the value to the nearest 5%, thus we generated a set of phenotype labels by rounding LVEF values to the nearest multiple of 5. For the final comparison, we generated binary LVEF labels by categorizing values as normal or abnormal, with normal values ranging from 52-72 for males and 54-74 for females.

3.3 Genome-wide association study

We used the UKB imputed genotype calls in BGEN v1.2 format. Samples were genotyped using the UK BiLEVE or UK Biobank Axiom arrays. Imputation was performed using the Haplotype Reference Consortium panel and the UK10K+1000 Genomes panel[12]**.** We used the QC files provided by UKB to create a GWAS cohort consisting of subjects who did not withdraw, were of inferred European ancestry, and were unrelated. Subjects with a genotype call rate < 0.98 were also removed. We considered variants with a minor allele frequency (MAF) ≥ 0.01, and we required genotyped variants to have a call rate ≥ 0.95 and imputed variants to have an INFO score ≥ 0.3. Variants with a Hardy-Weinberg equilibrium P value < $1x10^{-20}$ were excluded. After variant filtering, we were left with 9774199 filtered variants. GWAS was done on a Spark 3.1.1

cluster, using the library Hail 0.2 with Python version 3.6. The GWAS was adjusted for age at MRI and sex. We used the conventional P value of $5x10^{-8}$ as the threshold for defining genome-wide significance.

3.4 Assessing Association Power's Relationship with Cohort Size

Apart from noise in phenotype measurements, we also evaluate the effect of cohort decrease on GWAS results. We generated 6 different phenotype files where, starting from the original LVEF cohort (39,624), we keep 90% (35,661), 80% (31,699), 70% (27,736), 60% (23,774), 50% (19812), and 40% (15850) of the samples. Cohort decrease was performed before GWAS QC, and for each step the selection of samples to be excluded was random. Inspecting the effect of cohort decrease helps us define the relationship between the number of LVEF samples and GWAS power.

3.5 SNP-based accuracy

We use an accuracy metric to determine the amount of overlap in significant SNPs between the baseline GWAS results and noise-modified GWAS results. First, we remove all non-significant SNPs by excluding SNPs with a p-value less than 5×10^{-8}, which is the Bonferroni corrected p-value threshold. Then, we consider significant SNPs found in both the base results and noise-modified results as true positives (TP), the SNPs found only in the noise-modified results as false positives (FP), and the SNPs not found in the noise-modified results but found in the base results as false negatives (FN). We then calculate

$$SNP_{accuracy} = \frac{TP}{TP+FP+FN} \quad (2)$$

3.6 GWAS Sensitivity

Sensitivity determines the amount of overlap in significant loci between the baseline GWAS results and noise-modified GWAS results. Specifically, given that $peaks_{base}$ is the number of significant loci in base GWAS, and $peaks_{correct}$ is the number of significant loci that persisted in noise GWAS then

$$Sensitivity = \frac{peaks_{correct}}{peaks_{base}} \quad (3)$$

The number of loci and their position can be determined by manual inspection, but we also developed an automatic method. Our automatic method applies a hierarchical clustering algorithm on SNPs above the significance threshold line to determine the number and the position of loci from both GWAS, which we then use to compute $peaks_{base}$ and $peaks_{correct}$.

3.7 Heritability

Heritability is a measure of the level of influence genetic variation has on a given trait's phenotypic variation. To estimate SNP heritability based on GWAS summary statistic we use command line tool LDSC[16]. LDSC performs LD score regression between GWAS test statistic χ_j^2 and per SNP LD scored which allows for the estimation of h_g^2

4 Results

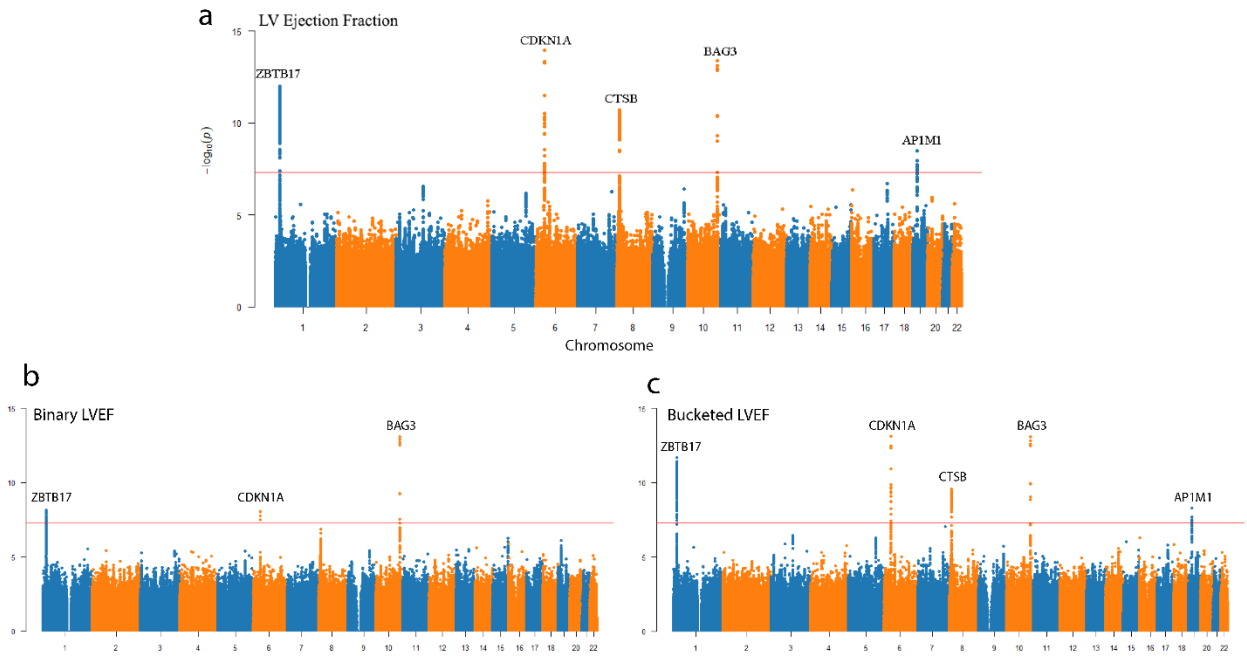

Figure 1 Manhattan plots for genome-wide association studies on UK Biobank reported left ventricular ejection fraction **a,** GWAS on continuous LVEF measurements **b,** GWAS on Normal/Abnormal LVEF where the range for normal is 52-72 in male and 54-74 female population **c,** GWAS on LVEF bucketed to the nearest multiple of 5

4.1 Quantitative phenotypes improve power of association studies

The study cohort for all analyses consisted of 39,624 adult unrelated subjects of European ancestry (**Table 1**). As a baseline, we first conducted a GWAS of the LVEF phenotype released with the UKBB cardiac MRI data. We identified 5 loci at genome-wide significance on chromosomes 1, 6, 8, 10, and 19 near genes *ZBTB17*, *CDKN1A*, *CTSB*, *BAG3*, and *AP1M1* (**Figure 1)**. In comparison, for an LVEF phenotype binarized to simply abnormal or normal, multiple previously detected loci lost genome-wide significance (including loci for *CTSB* and *AP1M1*). Similarly, recognizing the inherent variation present in measuring LVEF, we additionally compared the results if the LVEF was bucketed to 5% bins and showed such imprecision decreased statistical power in all SNPs in the association study compared to the continuous LVEF baseline phenotype.

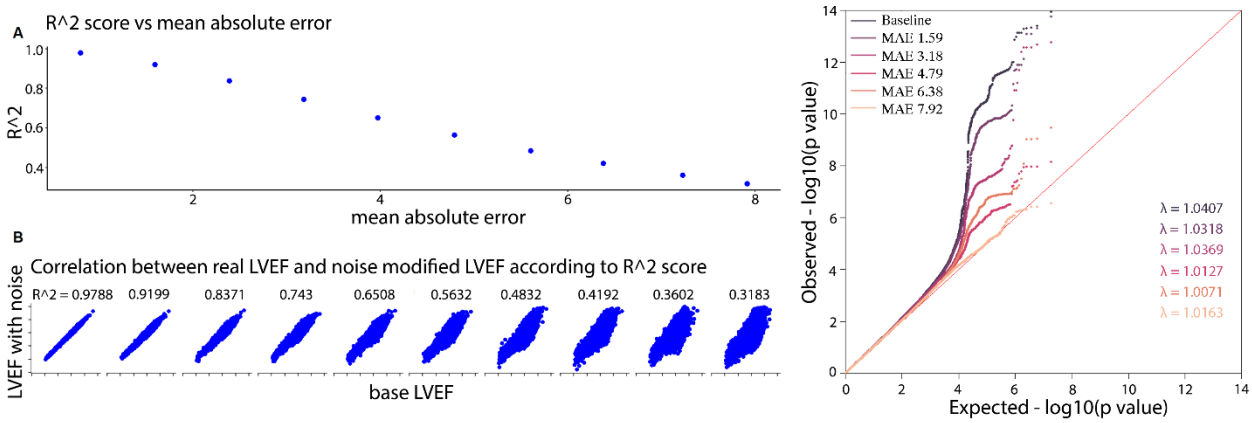

Figure 2 Impact of noise in LVEF on GWAS **a,** Visualizing r2 score, mean absolute error, and the distribution of noise-modified-LVEF with respect to the baseline LVEF. **b**, Q-Q plots of P values from GWAS summary statistics for different levels of noise

4.2 Phenotype noise degrades power of association studies

To investigate the effect of measurement imprecision on GWAS power, we performed a series of association studies while introducing noise in the range of known clinician variation **(Figure 2)**. Simulated variation to the LVEF measurement naturally increases in mean absolute error. Noise with a gaussian standard deviation of 5 results in a mean absolute error of 3.97% and R2 of 0.65 (**Table A1)**, and results in the loss of genome-wide significance for the *AP1M1* loci on chromosome 19. As we increase phenotypic noise in the range of clinical variation, heritability and power gradually declines and the noise equivalent to 7.92% MAE results in a complete loss of genomic-wide significance (**Table 2**). Given echocardiography is known to have a clinician-to-clinician variation of the same or greater MAE[10], such measurement imprecision could contribute to the limited hits in historial echocardiography-derived GWAS[8].

Table 2. Metrics of genetic signal for each increase in SD

Noise SD	SNP Accuracy	Loci Sensitivity	Heritability
0%	1.0	1.0	0.1114 (00357)
1%	0.9377	1.0	0.1055 (0.0332)
2%	0.8547	1.0	0.0878 (0.0352)
3%	0.3675	1.0	0.1003 (0.0265)
4%	0.2537	0.8	0.089 (0.0256)
5%	0.3921	0.8	0.1208 (0.0355)
6%	0.0228	0.4	0.0179 (0.0271)
7%	0.0307	0.4	0.0482 (0.0247)
8%	0.0145	0.4	0.022 (0.0349)
9%	0.0020	0.2	0.0355 (0.0204)
10%	0	0	0.0477 (0.0214)

4.3 Comparison of Impact of Phenotype Noise vs Cohort Size

Given the summary statistics from 16 different GWAS, we modeled the relationship between noise and GWAS power (**Figure 3, Figure A4)**. There is a linear relationship between the increase in MAE and the decrease in GWAS power. We calculated that an increase of 1% in MAE causes the loci sensitivity to decrease by 13% (p=5.5e-6) and the SNP accuracy by 14% (p=6.6e-5). Experiments with other methods of introducing noise in assessing LVEF similarly show a decrease in genetic association with more imprecise measurements **(Figure A3, Table A2)**. A similar effect occurs with reductions in cohort size, as a 1% decrease in cohort size results in a 1.3% decrease in loci sensitivity (p=0.01) and a 1.9% decrease in SNP-based accuracy (p=0.0007). We found that a 1% MAE increase has the same effect on loci sensitivity as a 10.3% cohort decrease and the same effect as a 7.2% cohort decrease on SNP accuracy.

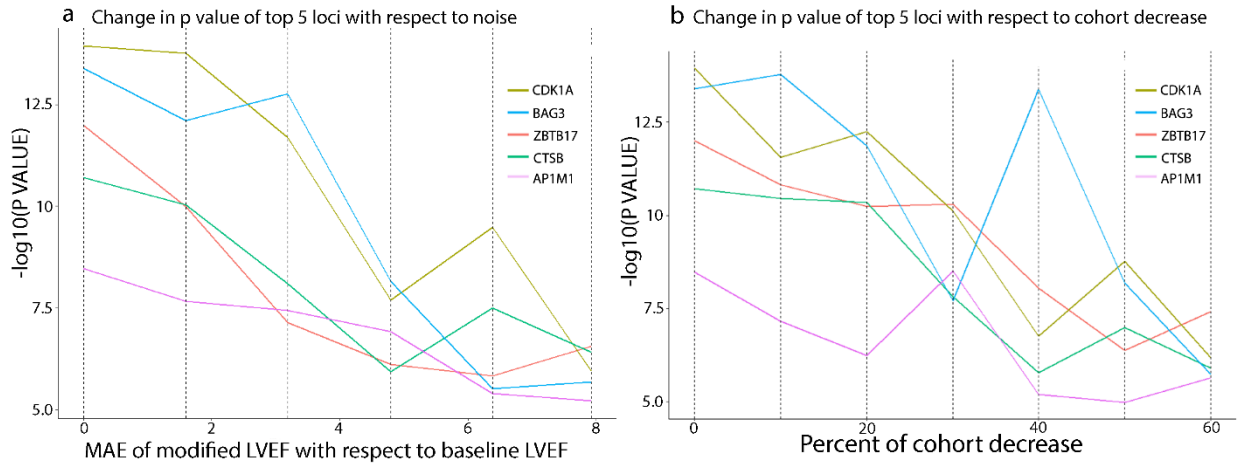

Figure 3 a, Slope chart shows the change in the P value of the top 5 loci with respect to mean absolute error; **b,** Slope chart shows the change in a P value of top 5 loci with respect to the cohort decrease; each locus is named after the closest gene

4.4 Improving phenotyping augments downstream genetic analyses

Cardiac MRI provides clinicians and researchers with a plethora of high-resolution imaging, with even the abbreviated 20-min UK Biobank cardiac MRI protocol resulting in 9 sequences with over 30,000 images per study[17]. With so many images and patients, the released UKBB measurements were generated using a fully automated workflow (with Siemens inLineVF) without quality inspection and bias correction. When compared with manual clinician evaluation, the automated measurements of LVEF result in a mean absolute error (MAE) of 3.4%, R2 of 0.348, and ICC of 0.521 for LVEF[15]. Imprecision in the inline LVEF can be partially addressed by linear adjustment[18] and doing so slightly increases genetic signal, within the difference in identified loci with MAE of 1% **(Figure A2)**. To evaluate the role of imprecision, we applied a deep learning-based method of obtaining LVEF and analyzed downstream results. Using a previously published deep learning segmentation model[6], we independently derived LV segmentation-based calculated LVEF and found a MAE of 6.1%, R2 of 0.335, and ICC of 0.431

for LVEF compared to the automated measurements from UKBB, and MAE of 5.3%, R2 of 0.60, and ICC of 0.518 compared to the linearly adjusted LVEF (**Figure 5)**. However, with these deep learning segmentation derived LVEF measurements, the same cohort identified more loci of interest with significant experimental data backing its relevance. In particular, loci on chromosomes 2, 5, and 8 near genes TTN, DNAJC18, and ZNF572 were not previously identified using the released UKBB LVEF measurements but able to be picked up with our quality-controlled measurements. While we could not directly compare the segmentation-derived LVEF measurements to clinical labels due to the absence of manual labels, the stronger genetic signal and higher association with linearly adjusted LVEF suggest that deep learning derived LVEF is less noisy.

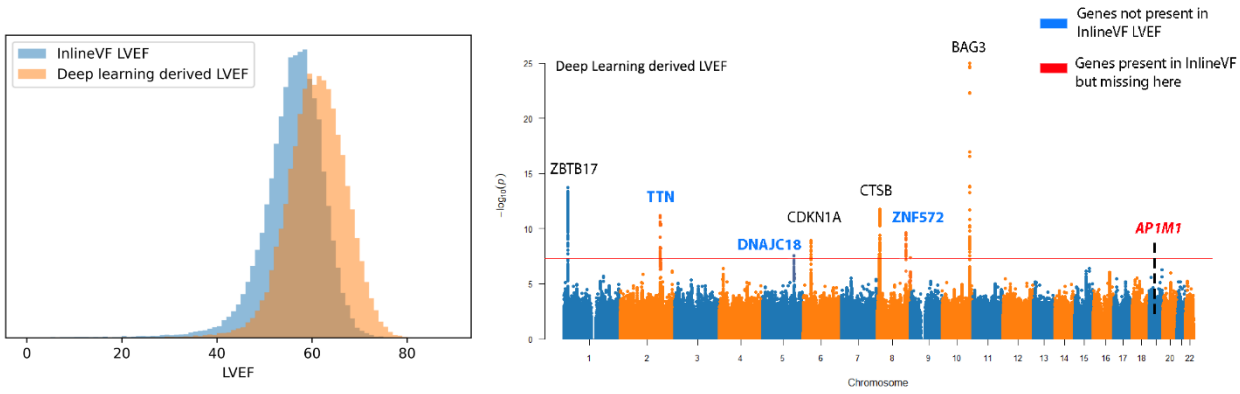

Figure 5 Differences in distribution and GWAS summary statistics between two methods of obtaining LVEF from MRI **a,** Histograms of InlineVF derived LVEF and Deep Learning derived LVEF **b,** Manhattan plot from GWAS performed on Deep Learning derived LVEF; genes colored in blue don't appear in InlineVF LVEF GWAS (Figure 1a); genes colored in red appear in InlineVF LVEF GWAS but not in deep learning derived LVEF GWAS

5 Discussion

In this study, we assessed the impact of measurement noise on genetic associations with LVEF and found substantially impaired power in downstream GWAS analysis with even slight increases in measurement imprecision. Even slight phenotyping variation can significantly impact downstream genetic associations, often to a greater extent than changes in cohort size. As measurement variation is present in many clinical measurements, efforts to improve the precision of measurements can potentially be a cost-effective way to maximize the yield of genetic association studies.

Cardiac function as measured by LVEF is an important clinical measurement that defines disease and identifies patients who are eligible for life-prolonging therapeutics as implantable devices. In echocardiography, human test-retest evaluation of LVEF can range between 7-10%, with slight changes in annotation as well as timing that can significantly impact calculations[10,19]. Few variability studies have been undertaken in cardiac MRI, although similar degrees of manual measurement variability have been found[20]. Prior studies have suggested that polygenic risk

scores of LVESV have more power than polygenic risk scores of LVEF[2], consistent with our analyses that more precise measurements correspond to stronger genetic associations. Our analysis suggests that a substantial and primary gain in signal comes from the improvement of noisy measurements that can affect the power and accuracy of downstream analyses.

Noise in measurements can appear in both semi-automated and fully automated workflows[11], and by improving the precision of measuring LVEF, we also improve the accuracy and robustness of downstream GWAS results. The relatively large improvement in yield of genetic association with more precise phenotyping was substantial in comparison to the marginal benefit of increasing the cohort size. As more genetic analyses are undertaken with automated measurements or assessments[4,7,21,22], an additional evaluation must be taken to assess the variability and quality of the phenotyping. Such insights ideally will be confirmed with orthogonal measurements of similar phenotypes. Some of the first genetic association studies were performed on quantitative traits like height, but it should be recognized that many imaging-based phenotypes do not have the same precision and accuracy as the assessment of height on a population.

In summary, genetic association studies on imaging phenotypes allow researchers to discover many associations that help understand the underlying biology of the disease and structure[23]. For LVEF, even advanced imaging has variability in measurements that can substantially impact downstream association studies. The impact of such variability is even more profound than significant changes in cohort size, suggesting improvement in imaging precision and precise phenotyping in general has significant additional value in improving the power of genetic association studies.

Our study offers key insights into measurement noise's effect on genetic associations with LVEF. However, a few considerations remain. The impact of measurement noise could vary for different quantitative phenotypes, and thus future studies should investigate its influence on various phenotypes for a broader understanding. Secondly, our GWAS methodology could be further enhanced by using a linear mixed model method[24], shown to produce more significant associations. Lastly, while our deep learning LVEF method showed a high GWAS signal, we could not compare it to manual clinical labels due to their unavailability.

6 Appendix

Table A1. Mapping between Gaussian Noise SD and MAE

SD	MAE	R2
0	0	1
1	0.797489	0.9788
2	1.594416	0.9199
3	2.386753	0.8371
4	3.183924	0.743
5	3.974958	0.6508
6	4.793956	0.5632
7	5.604129	0.4832
8	6.380848	0.4192
9	7.228321	0.3602
10	7.920860	0.3183

Table A2. Metrics of genetic signal for each decrease in cohort size

Cohort decrease	SNP Accuracy	GWAS Sensitivity	Heritability
0%	1.0	1.0	0.1114 (00357)
10%	0.8744	0.8	0.1071 (0.0397)
20%	0.8713	0.8	0.0867 (0.037)
30%	0.3436	1.0	0.082 (0.0332)
40%	0.1392	0.4	0.0497 (0.0216)
50%	0.0477	0.4	0.039 (0.0287)
60%	0.0019	0.2	0.0384 (0.0288)

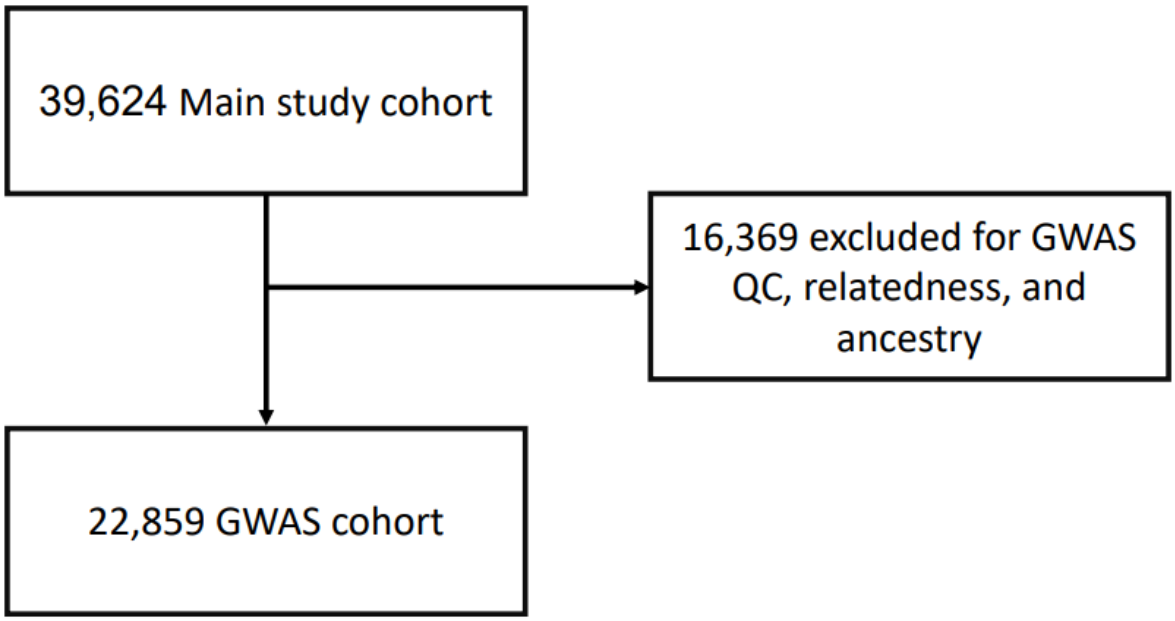

Figure A1. Cohort diagram

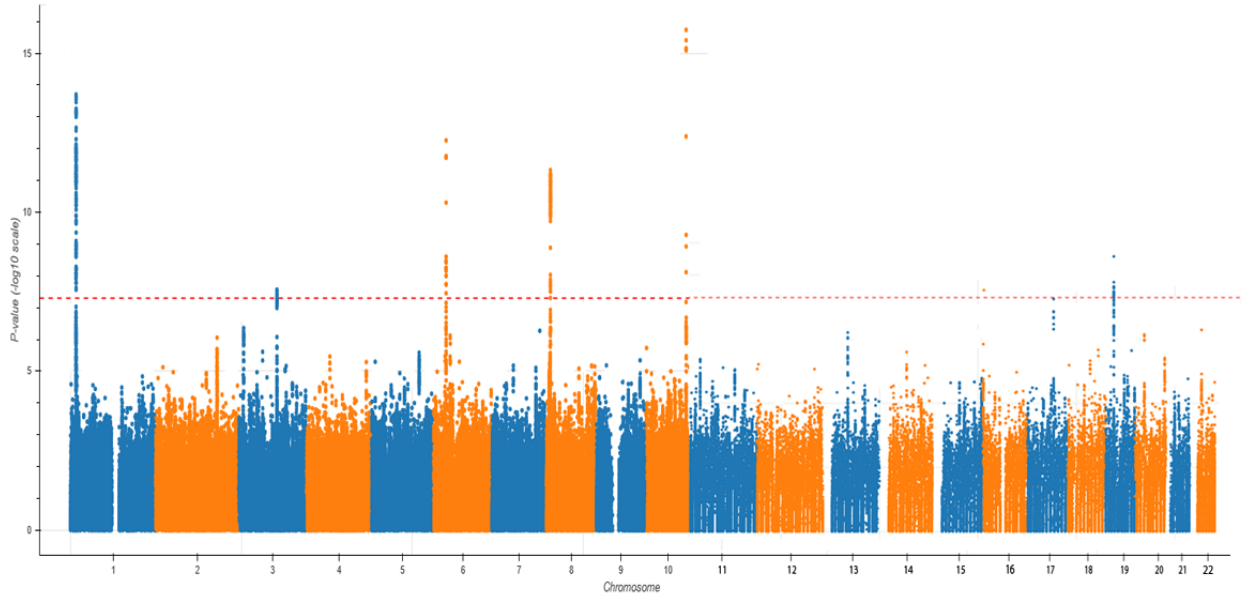

Figure A2. Manhattan plot for genome-wide association study on corrected left-ventricular ejection fraction.

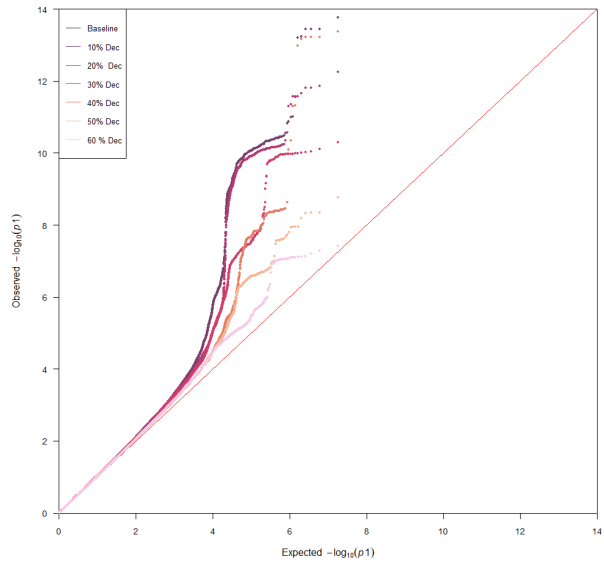

Figure A3. Q-Q plots of P values from GWAS summary statistics for different percentages of cohort decrease

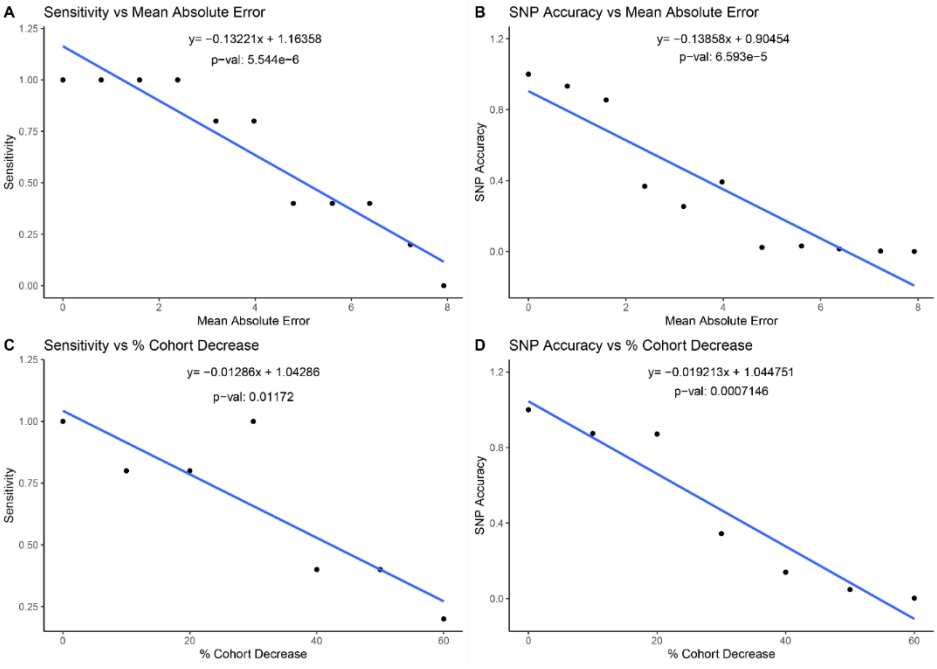

Figure A4 Impact of cohort decrease and noise generation on GWAS power. **a**, Regression analysis on the impact of measurement error quantified by a mean absolute error on sensitivity. **b**, Regression analysis on the impact of the mean absolute error on SNP accuracy. **c**, Regression analysis of the impact of cohort size decline on sensitivity. **d**, Regression analysis of the impact of cohort size decline on SNP accuracy

References

1. Pirruccello, J. P. *et al.* Genetic analysis of right heart structure and function in 40,000 people. *bioRxiv* (2021) doi:10.1101/2021.02.05.429046.
2. Pirruccello, J. P. *et al.* Analysis of cardiac magnetic resonance imaging in 36,000 individuals yields genetic insights into dilated cardiomyopathy. *Nat. Commun.* **11**, 2254 (2020).
3. Agrawal, S. *et al.* Inherited basis of visceral, abdominal subcutaneous and gluteofemoral fat depots. *Nat. Commun.* **13**, 3771 (2022).
4. Haas, M. E. *et al.* Machine learning enables new insights into genetic contributions to liver fat accumulation. *Cell Genom* **1**, (2021).
5. Liu, D. J. *et al.* Exome-wide association study of plasma lipids in >300,000 individuals. *Nat. Genet.* **49**, 1758–1766 (2017).
6. Bai, W. *et al.* Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *J. Cardiovasc. Magn. Reson.* **20**, 65 (2018).
7. Meyer, H. V. *et al.* Genetic and functional insights into the fractal structure of the heart. *Nature* **584**, 589–594 (2020).
8. Vasan, R. S. *et al.* Genetic variants associated with cardiac structure and function: a meta-analysis and replication of genome-wide association data. *JAMA* **302**, 168–178 (2009).
9. Carroll, R. J. *et al.* Nonparametric Prediction in Measurement Error Models [with Commants]. *J. Am. Stat. Assoc.* **104**, 993–1014 (2009).
10. Farsalinos, K. E. *et al.* Head-to-Head Comparison of Global Longitudinal Strain Measurements among Nine Different Vendors: The EACVI/ASE Inter-Vendor Comparison Study. *J. Am. Soc. Echocardiogr.* **28**, 1171–1181, e2 (2015).
11. O'Dell, W. G. Accuracy of Left Ventricular Cavity Volume and Ejection Fraction for Conventional Estimation Methods and 3D Surface Fitting. *J. Am. Heart Assoc.* **8**, e009124 (2019).
12. Littlejohns, T. J. *et al.* The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nat. Commun.* **11**, 2624 (2020).
13. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
14. Petersen, S. E. *et al.* UK Biobank's cardiovascular magnetic resonance protocol. *J. Cardiovasc. Magn. Reson.* **18**, 8 (2016).
15. Suinesiaputra, A. *et al.* Fully-automated left ventricular mass and volume MRI analysis in the UK Biobank population cohort: evaluation of initial results. *Int. J. Cardiovasc. Imaging* **34**, 281–291 (2018).
16. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
17. Petersen, S. E. *et al.* UK Biobank's cardiovascular magnetic resonance protocol. *J. Cardiovasc. Magn. Reson.* **18**, 1–7 (2016).
18. Sanghvi, M. M. *et al.* Automatic left ventricular analysis with Inline VF performs well compared to manual analysis: results from Barts Cardiovascular Registry. *J. Cardiovasc. Magn. Reson.* **18**, 1–2 (2016).

19. Yuan, N. *et al.* Systematic Quantification of Sources of Variation in Ejection Fraction Calculation Using Deep Learning. *JACC Cardiovasc. Imaging* **14**, 2260–2262 (2021).
20. Augusto, J. B. *et al.* Diagnosis and risk stratification in hypertrophic cardiomyopathy using machine learning wall thickness measurement: a comparison with human test-retest performance. *Lancet Digit Health* **3**, e20–e28 (2021).
21. Zekavat, S. M. *et al.* Deep Learning of the Retina Enables Phenome- and Genome-Wide Analyses of the Microvasculature. *Circulation* **145**, 134–150 (2022).
22. Kosaraju, A., Goyal, A., Grigorova, Y. & Makaryus, A. N. Left Ventricular Ejection Fraction. in *StatPearls* (StatPearls Publishing, 2022).
23. Uffelmann, E. *et al.* Genome-wide association studies. *Nature Reviews Methods Primers* **1**, 1–21 (2021).
24. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).