An Empirical Study of ML-based Phenotyping and Denoising for Improved Genomic Discovery

Bo Yuan*, Cory McLean, Farhad Hormozdiari, Justin Cosentino*
*boyuan@g.harvard.edu, jtcosentino@google.com
Genomics team in Health AI

Google Research

Introduction

Goal

Empirically evaluate the robustness of ML-based phenotyping to varying levels of label noise.

Motivation

ML-based phenotyping, in which an ML model is applied to high dimensional clinical data to predict a target phenotype, enables fast and accurate phenotyping at biobank scales. Though recent work demonstrates the application of this method to low quality labels, it has not been possible to quantify changes in genetic association power since the underlying ground-truth liability scores for complex, polygenic diseases remain unknown. To address this challenge, we corrupt a continuous phenotype using varying levels of noise and study changes in discovery.

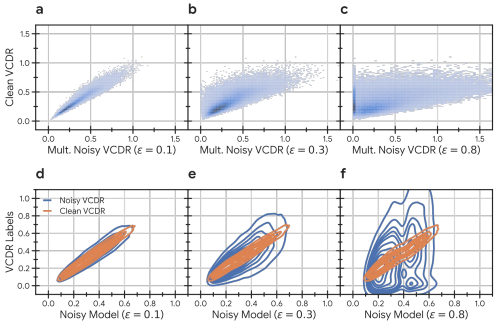

Figure 1: ML-based phenotyping is robust label corruption.

Highlights

- We simulate label corruption by applying varying levels of random noise to vertical cup-to-disc ratio (VCDR).
- We show that the standard ML-based phenotyping procedure is reasonably robust across noise levels.
- We propose an integrated denoising approach to the ML-based phenotyping procedure.
- We evaluate the impact of noise on downstream genomic discovery and polygenic risk score performance.

Method

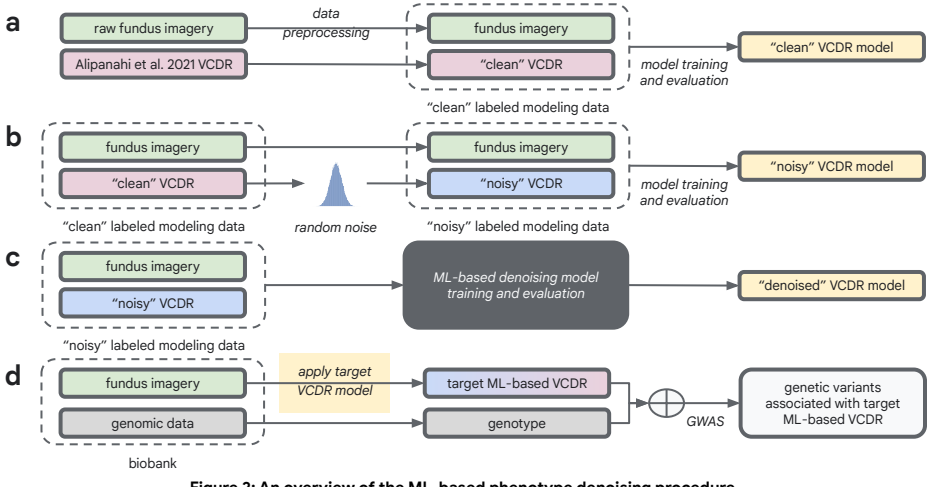

Figure 2: An overview of the ML-based phenotype denoising procedure.

The training process for the a) *clean*, b) *noisy*, and c) *denoised* VCDR models. We d) applied each model to UK Biobank fundus imagery and used the liability scores for genomic discovery.

Results

Phenotype	ML Noisy R	ML Clean R	NCP	Replication %	PRS Euro R	PRS Non-Euro R
Clean VCDR	-	1.0000 ± 0.0000	194.7313 ± 0.0000	100%	0.3047 ± 0.0230	0.2800 ± 0.0096
Noisy VCDR ($\epsilon = 0.1$)	-	0.9685 ± 0.0005	181.7726 ± 0.0999	92.86%	0.1738 ± 0.0239	0.1509 ± 0.0096
Noisy Model	0.9511 ± 0.0008	0.9819 ± 0.0004	185.8213 ± 0.0923	96.75%	0.3974 ± 0.0223	0.3652 ± 0.0096
Denoised Model (binary)	0.9506 ± 0.0008	0.9819 ± 0.0003	185.7674 ± 0.0794	96.10%	0.4083 ± 0.0221	0.3712 ± 0.0092
Denoised Model (oracle)	0.9484 ± 0.0008	0.9795 ± 0.0004	184.8610 ± 0.1024	96.75%	0.4008 ± 0.0225	0.3743 ± 0.0094
Noisy VCDR ($\epsilon = 0.3$)	-	0.7941 ± 0.0035	118.6192 ± 0.4719	76.30%	0.1347 ± 0.0240	0.1188 ± 0.0095
Noisy Model	0.7677 ± 0.0036	0.9664 ± 0.0006	178.2096 ± 0.1386	93.51%	0.3904 ± 0.0222	0.3539 ± 0.0093
Denoised Model (binary)	0.7669 ± 0.0038	0.9661 ± 0.0007	178.4808 ± 0.1500	94.16%	0.3904 ± 0.0231	0.3607 ± 0.0096
Denoised Model (oracle)	0.7713 ± 0.0037	0.9729 ± 0.0005	181.6247 ± 0.1259	95.13%	0.3933 ± 0.0219	0.3645 ± 0.0092
Noisy VCDR ($\epsilon = 0.8$)	-	0.4932 ± 0.0070	43.0540 ± 0.5012	41.88%	0.0975 ± 0.0261	0.1451 ± 0.0105
Noisy Model	0.4652 ± 0.0069	0.9443 ± 0.0009	168.1538 ± 0.2195	91.56%	0.4009 ± 0.0217	0.3596 ± 0.0094
Denoised Model (binary)	0.4610 ± 0.0071	0.9373 ± 0.0011	164.3056 ± 0.2331	89.94%	0.3948 ± 0.0232	0.3719 ± 0.0094
Denoised Model (oracle)	0.4642 ± 0.0071	0.9463 ± 0.0011	168.9465 ± 0.2057	94.16%	0.3845 ± 0.0220	0.3574 ± 0.0096

Table 1: ML-based phenotyping recovers the underlying liability score across noise levels, significantly improving genetic discovery and PRS predictive power relative to noisy equivalents.

ML Noisy R and *ML Clean R* denote Pearson's correlation between labels or model predictions and the target noisy or ground-truth labels. *NCP* denotes the non-centrality parameter, a proxy for GWAS power. *Replication %* captures the percent of ground-truth GWS hits replicated. *PRS Euro R* and *PRS Non-Euro R* denote the correlation between PRS scores in the European holdout set (n=1,472) and the non-European validation set (n=10,095).

Conclusion

Takeaways

- Standard ML-based phenotyping approaches successfully recover underlying liability scores given corrupted labels.
- ML-based phenotyping significantly improves PRS predictive power relative to both the ground-truth and noisy GWAS.
- Our SNVC-based denoising method shows promising initial results for integrated approaches.

Future Directions

- Extending this analysis to the binary label setting to better mirror the nature of the EHRs often found in biobanks
- Evaluating other noise distributions (e.g., structured noise) to better understand the impact of systematic dataset bias
- Further improving integrated denoising methods

Resources

bioRxiv preprint: biorxiv.org/content/10.1101/2022.11.17.516907v1
Open source code: github.com/Google-Health/genomics-research/tree/main/ml-based-vcdr

References

Alipanahi et al., "Large-scale machine-learning-based phenotyping significantly improves genomic discovery for optic nerve head morphology," AJHG 2021.
Cosentino et al., "Leveraging deep-learning on raw spirograms to improve genetic understanding and risk scoring of COPD despite noisy labels," bioRxiv 2022.