Validating Orchid’s Celiac Disease Genetic Risk Score

Written by Orchid Bioinformatics Staff

Introduction

Celiac disease is a chronic autoimmune disorder in which ingestion of gluten triggers an immune response that damages the lining of the small intestine, leading to impaired nutrient absorption. Common symptoms include diarrhea, fatigue, weight loss, bloating, and anemia.[1] The prevalence of celiac disease in the U.S. general population is estimated to be approximately 1%.[2] Treatment requires lifelong adherence to a strict gluten-free diet, which typically leads to symptom resolution and intestinal healing. Although there is no cure, early diagnosis and proper dietary management significantly reduce the risk of long-term complications.[1]

Genetic Risk Score

A person’s risk of developing celiac disease is shaped substantially by genetics.[3] Genetic risk scores (GRS) allow us to estimate this disease risk based on the DNA of a person or embryo.[4] Although not diagnostic, a GRS can indicate how likely an individual is to develop the disease compared to the population baseline risk.

The celiac disease GRS used in Orchid’s reports includes 82 variants and was developed based on a study that included 12,041 cases (individuals with celiac disease) and 12,228 healthy controls.[5,6] The final GRS score gives special weight to specific variants within the immune-related HLA region of chromosome 6, which contribute a disproportionately large share of genetic risk.[7]

Genetic risk scores that were trained primarily on data from people of European ancestry tend to be less accurate when applied to individuals of non-European ancestry.[8] This problem is especially challenging in the case of celiac disease, and so we unfortunately only offer the celiac disease model predictions to individuals of European ancestry at this time, although research efforts are currently underway to improve the predictive accuracy of this model for other ancestry groups.

Evaluation on UK Biobank Data

We evaluated the predictive accuracy of Orchid’s celiac disease GRS using the UK Biobank (UKB), a research database of roughly 500,000 genotyped individuals from the United Kingdom.[9] We restricted the analysis to people of British ancestry and defined celiac disease using the K90.0 ICD-10 code. This yielded 2,708 cases and 405,812 controls (0.66% prevalence). We then grouped individuals by GRS percentile and compared the observed disease prevalence within each group to our model’s predictions (Figure 1). For additional technical details, see the Supplementary Information.

Table 1 shows the celiac disease observed prevalence for individuals in the UKB grouped by GRS percentile range (top 10%, 5%, and 1%), as well as how their risk compares to the baseline risk at the 50th GRS percentile. Those with higher GRS relative to the population baseline also had substantially higher observed prevalence of celiac disease, supporting the predictive accuracy of the GRS to identify individuals with elevated risk.

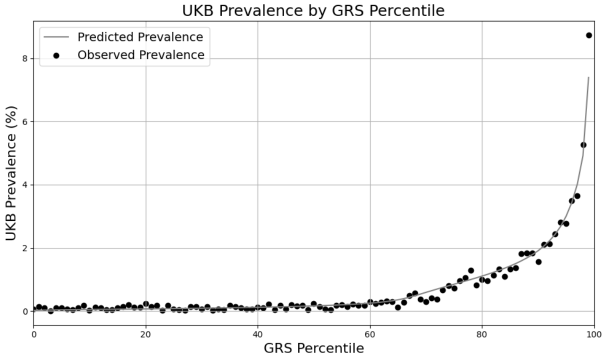

Figure 1. Risk Stratification. Predicted and observed prevalence in the UKB for individuals grouped by GRS percentile.

GRS Group	Observed UKB Prevalence	Odds Ratio
Baseline (50th percentile)	0.15%	1.00
Top 10%	3.50%	24.64
Top 5%	4.78%	34.15
Top 1%	8.74%	65.08

Table 1. Observed prevalence of celiac disease in the UKB by GRS percentile range. Those with higher GRS relative to the population baseline also had substantially higher observed prevalence of celiac disease.

Estimating Lifetime Risk

The average observed prevalence of celiac disease in the UKB was 0.66%. This is lower than the lifetime prevalence in the general population, which has been estimated to be approximately 1%.[2] This is likely due in part to the fact that UKB participants tend to be healthier than the general population, which leads to lower observed disease prevalence.[10] Additionally, the observed prevalence in the UKB includes people still living who could develop the disease when they are older, and so does not capture the full lifetime risk of the disease.

Orchid's clinical reports include predicted lifetime disease risk, which we calculate by first estimating how disease risk varies across GRS in the UKB and then rescaling that pattern so the average matches the known lifetime population risk (Figure 2).[4] People at the high end of the GRS distribution are predicted to have an elevated lifetime risk of the disease relative to the population (Table 2).

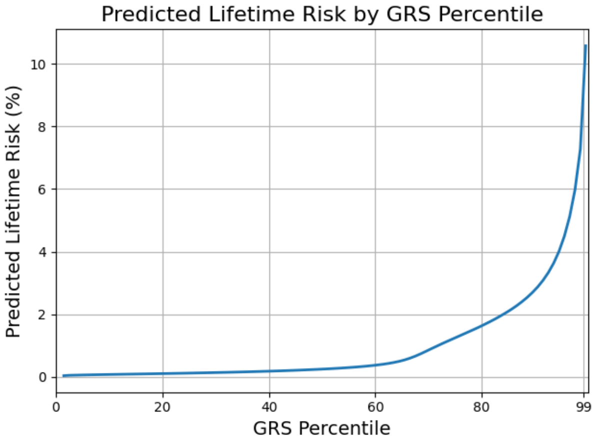

Figure 2. Adjusted Risk Stratification. Predicted risk estimates adjusted so that overall prevalence matches the approximately 1% estimate.[2]

GRS Percentile	Predicted Lifetime Risk	Relative Risk
50th (baseline)	0.25%	1.00x
95th	4.23%	17.14x
97th	5.51%	22.32x
99th	8.25%	33.44x

Table 2. Predicted lifetime prevalence of celiac disease at different GRS percentiles. Individuals with the highest GRS percentiles are predicted to have an increased risk of celiac disease relative to those at the 50th percentile.

Conclusion

In this study, we evaluated our celiac disease GRS on data from the UKB. We found that it performed well, particularly for identifying individuals with elevated risk of the disease relative to the population. In our embryo and couple reports, we adjust the model to predict lifetime risk, which is generally higher than observed prevalence in the UKB. The celiac disease GRS model is currently only available to individuals of European ancestry, due to concerns about loss of accuracy in other ancestry groups, although research efforts are underway to address this issue.

Acknowledgements

This research was conducted using the UK Biobank Resource under Application Number 80545.

References

1. Mayo Clinic Staff. Celiac Disease. https://www.mayoclinic.org/diseases-conditions/celiac-disease/symptoms-causes/syc-20352220, 2024. Accessed 2026.
2. GK Makharia, A Chauhan, P Singh, and V Ahuja. Review Article: Epidemiology of Coeliac Disease. *Alimentary Pharmacology & Therapeutics*, 56(Suppl 1):S3–S17, 2022. doi:10.1111/apt.16787.
3. R Kuja-Halkola, B Lebwohl, J Halfvarson, C Wijmenga, PK Magnusson, and JF Ludvigsson. Heritability of Non-HLA Genetics in Coeliac Disease: A Population-Based Study in 107,000 Twins. *Gut*, 65(11):1793–1798, 2016. doi:10.1136/gutjnl-2016-311713.
4. N Chatterjee, J Shi, M García-Closas, et al. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics*, 17:392–406, 2016. doi:10.1038/nrg.2016.27.
5. SA Sharp, SE Jones, RA Kimmitt, et al. A Single Nucleotide Polymorphism Genetic Risk Score to Aid Diagnosis of Coeliac Disease: A Pilot Study in Clinical Care. *Alimentary Pharmacology & Therapeutics*, 52(7):1165–1173, 2020. doi:10.1111/apt.15826.
6. G Trynka, KA Hunt, NA Bockett, et al. Dense Genotyping Identifies and Localizes Multiple Common and Rare Variant Association Signals in Celiac Disease. *Nature Genetics*, 43(12):1193–1201, 2011. doi:10.1038/ng.998.
7. JA Murray, SB Moore, CT Van Dyke, et al. HLA DQ Gene Dosage and Risk and Severity of Celiac Disease. *Clinical Gastroenterology and Hepatology*, 5(12):1406–1412, 2007. doi:10.1016/j.cgh.2007.08.013.
8. Florian Privé, Hugues Aschard, Shai Carmi, et al. Portability of 245 Polygenic Scores When Derived from the UK Biobank and Applied to 9 Ancestry Groups from the Same Cohort. *American Journal of Human Genetics*, 109(1):12–23, 2022. doi:10.1016/j.ajhg.2021.11.008.
9. C. Sudlow, J. Gallacher, N. Allen, et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, 12(3):e1001779, 2015. doi:10.1371/journal.pmed.1001779.
10. A. Fry, T.J. Littlejohns, C. Sudlow, et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *American Journal of Epidemiology*, 186:1026–1034, 2017. doi:10.1093/aje/kwx246.

Supplementary Information

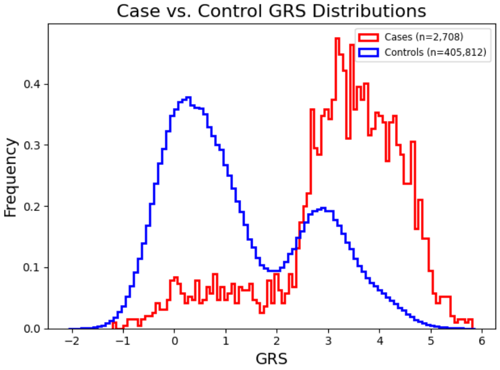

Figure 3. GRS histograms. GRS distributions for cases and controls. The case distribution is shifted noticeably higher compared to the controls.

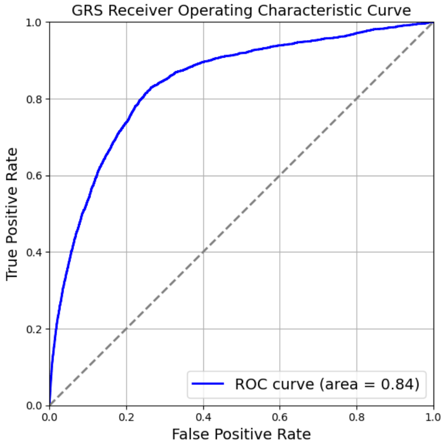

Figure 4. The receiver operating characteristic (ROC) used to compute the ROC area under the curve (AUC). The ROC curve is a graphical representation of a binary classifier's performance, plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) across different decision thresholds. A curve closer to the top-left indicates a better model, while a diagonal line (AUC = 0.5) represents random guessing.