Validating Orchid's Prostate Cancer Genetic Risk Score

ORCHID

Written by Orchid Bioinformatics Staff

Introduction

Prostate cancer is a disease in which malignant cells form in the prostate, a gland in the male reproductive system.[1] Common symptoms include difficulty urinating, a weak or interrupted urine stream, frequent urination (especially at night), pain or burning during urination, and persistent pain in the lower back, hips, or pelvis.[2] Risk for prostate cancer is influenced by both genetic and non-genetic factors. Non-genetic risk factors include increasing age, certain dietary factors, and exposure to certain chemicals such as arsenic.[3]

Prostate cancer affects roughly 3.5 million men in the United States, and approximately 12.9% will be diagnosed during their lifetime.[4] Prognosis for prostate cancer varies considerably by stage at diagnosis. The 5-year relative survival rate is more than 99% for localized disease (cancer confined to the prostate), more than 99% for regional disease, and about 38% for metastatic disease.[5] Many cases of prostate cancer can be managed with active surveillance and do not require immediate treatment, but some may require surgery, radiation therapy, or other medical treatments prescribed by an oncologist.[6]

Genetic Risk Score

Prostate cancer is shaped by both environmental and genetic factors. Monogenic testing is not available because no single gene causes the condition. Genetic risk scores (GRS), which combine the small effects of many variants into a single score, are currently the only way to estimate genetic risk. Although not diagnostic, a GRS can indicate how likely an individual is to develop the disease compared to the population baseline risk.

The prostate cancer GRS was developed using data from multiple large studies, including the PRACTICAL Consortium, UK Biobank (UKB), FinnGen, and All of Us, consisting of individuals of European ancestry.[7–10] For the UKB, a subset of individuals was held out to use in the validation study. Summary statistics were generated for the UKB and All of Us using the Regenie method.[11] Summary statistics across all studies were then meta-analyzed using METAL.[12] The final GRS was trained using SBayesRC and includes over 7 million variants.[13]

Risk predictions are adjusted to each individual's ancestry, with predictive power decaying as genetic distance from the predominantly European training data increases.[14] Orchid considers a GRS meaningfully predictive if individuals at approximately the 97.7th percentile have an odds ratio (OR) of at least 2. The prostate cancer GRS meets this criterion for all common ancestry groups.

Evaluation on UK Biobank Data

We evaluated the predictive accuracy of Orchid's prostate cancer GRS using the UK Biobank (UKB), a research database of roughly 500,000 genotyped individuals from the United Kingdom.[8] Because some data from the UKB was used to train the GRS, our validation study included only the set of samples that were held out for testing and not included in the GRS training. We also restricted the analysis to males of British ancestry and defined prostate cancer using the C61 ICD-10 code, yielding 2,909 cases and 46,299 controls (5.9% 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.

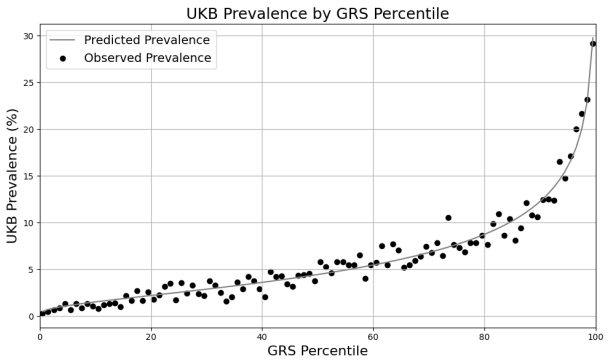

Figure 1. Risk Stratification. Predicted and observed prevalence in the UKB for individuals grouped by GRS percentile.

Table 1 shows the prostate cancer 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 prostate cancer, supporting the predictive accuracy of the GRS to identify individuals with elevated risk.

GRS Group	Observed UKB Prevalence	Odds Ratio
Baseline (50th percentile)	4.85%	1.00
Top 10%	17.99%	4.30
Top 5%	22.24%	5.61
Top 1%	29.14%	8.06

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

Estimating Lifetime Risk

The average observed prevalence of prostate cancer in the UKB was 5.9%. This is considerably lower than the lifetime prevalence in the US general population, which has been estimated to be approximately 12.9%.[4] 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.[15] 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).[16] 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).

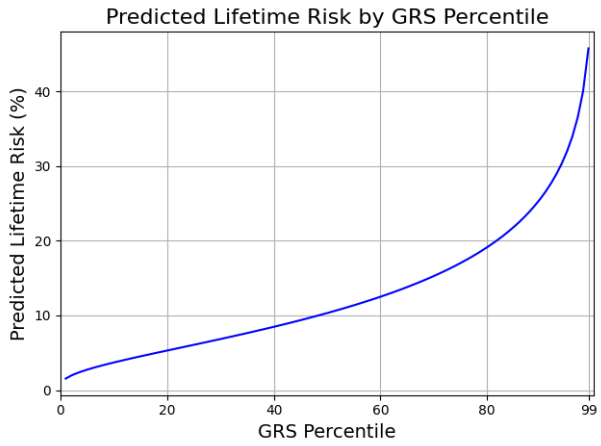

Figure 2. Adjusted Risk Stratification. Predicted risk estimates adjusted so that overall prevalence matches the 12.9% estimate from the National Cancer Institute.[4]

GRS Percentile	Predicted Lifetime Risk	Relative Risk
50th (baseline)	10.32%	1.00x
95th	32.02%	3.10x
97th	36.57%	3.54x
99th	45.78%	4.44x

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

Conclusion

In this study, we evaluated our prostate cancer 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 prostate cancer GRS model is available to individuals of all ancestry groups.

Acknowledgments

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

References

1. National Cancer Institute. Prostate cancer. In NCI Dictionary of Cancer Terms, 2025. URL: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/prostate-cancer.
2. Centers for Disease Control and Prevention. Symptoms of prostate cancer, February 11 2025. URL: https://www.cdc.gov/prostate-cancer/symptoms/index.html.
3. American Cancer Society. What causes prostate cancer?, November 22 2023. URL: https://www.cancer.org/cancer/types/prostate-cancer/causes-risks-prevention/what-causes.html.
4. U.S. National Cancer Institute. Prostate cancer — Cancer Stat Facts. SEER, 2025. URL: https://seer.cancer.gov/statfacts/html/prost.html.
5. American Cancer Society. What are the survival rates for prostate cancer?, 2025. URL: https://www.cancer.org/cancer/types/prostate-cancer/detection-diagnosis-staging/survival-rates.html.
6. American Cancer Society. Observation or active surveillance for prostate cancer, 2025. URL: https://www.cancer.org/cancer/types/prostate-cancer/treating/watchful-waiting.html.
7. Fredrick R. Schumacher, Ali A. Al Olama, Sonja I. Berndt, et al. Association analyses of more than

140,000 men identify 63 new prostate cancer susceptibility loci. *Nature Genetics*, 50(7):928–936, 2018. doi:10.1038/s41588-018-0142-8.

8. 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.
9. FinnGen Consortium. FinnGen documentation of R12 release. 2023. URL: https://finngen.gitbook.io/documentation/.
10. All of Us Research Program Investigators. The "All of Us" Research Program. *New England Journal of Medicine*, 381(7):668–676, 2019. doi:10.1056/NEJMsr1809937.
11. J. Mbatchou, L. Barnard, J. Backman, et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics*, 53:1097–1103, 2021. doi:10.1038/s41588-021-00870-7.
12. C.J. Willer, Y. Li, and G.R. Abecasis. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 26(17):2190–2191, 2010. doi:10.1093/bioinformatics/btq340.
13. Z. Zheng et al. Leveraging functional genomic annotations and LD structure to improve polygenic prediction. *Nature Communications*, 13:1–12, 2022. doi:10.1038/s41467-022-29849-5.
14. Florian Privé, Hugues Aschard, Andrey Ziyatdinov, and Michael G.B. Blum. 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.
15. 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.
16. 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.

Supplementary Information

Baseline Risk	OR per SD	OR per 2 SD
10.32%	2.36	5.55

Table 3. OR per SD. The baseline risk for an individual with a median GRS, and the predicted OR at one and two SDs, respectively. A GRS must have a predicted OR >2 at 2 SD to be included in Orchid's clinical reports.

UKB Prevalence	Population Prevalence	Liability R^2
5.9%	12.9%	18.11%

Table 4. Liability R^2. The estimated liability R^2 using a population prevalence of 12.9%.

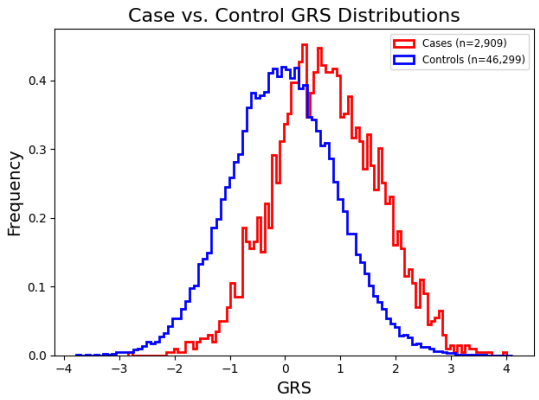

Figure 3. GRS histograms. GRS distributions for cases and controls. Both are approximately normal, with the case distribution shifted noticeably higher compared to the controls.

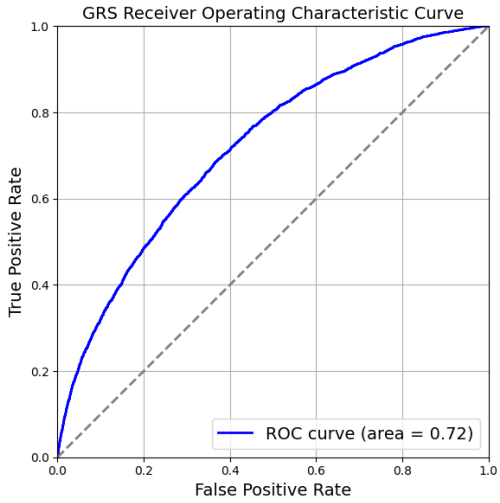

Figure 4. The receiver operating characteristic (ROC) used to compute the ROC area under the curve (AUC). The ROC curve is a graphical representation of a binary classifier's performance, plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) across different decision thresholds.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

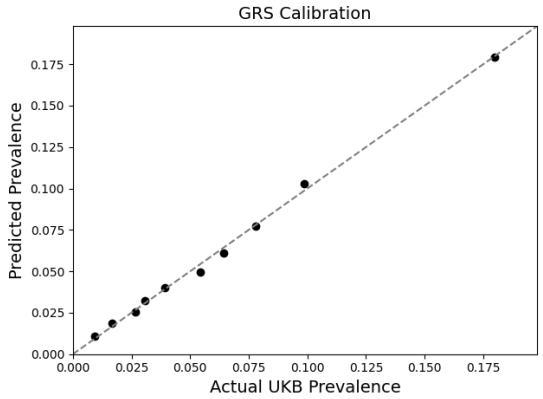

Figure 5. Calibration Curve. Calibration plot showing observed disease prevalence versus predicted risk across GRS deciles.