NPDR Supplementary Material

Trang T. Le[1], Bryan A. Dawkins[2] and Brett A. McKinney[2,3*]

[1]Department of Biostatistics, Epidemiology and Informatics,
University of Pennsylvania, Philadelphia, PA 19104
[2]Department of Mathematics, University of Tulsa, Tulsa, OK 74104
[3]Tandy School of Computer Science, University of Tulsa, Tulsa,
OK 74104

March 4, 2019

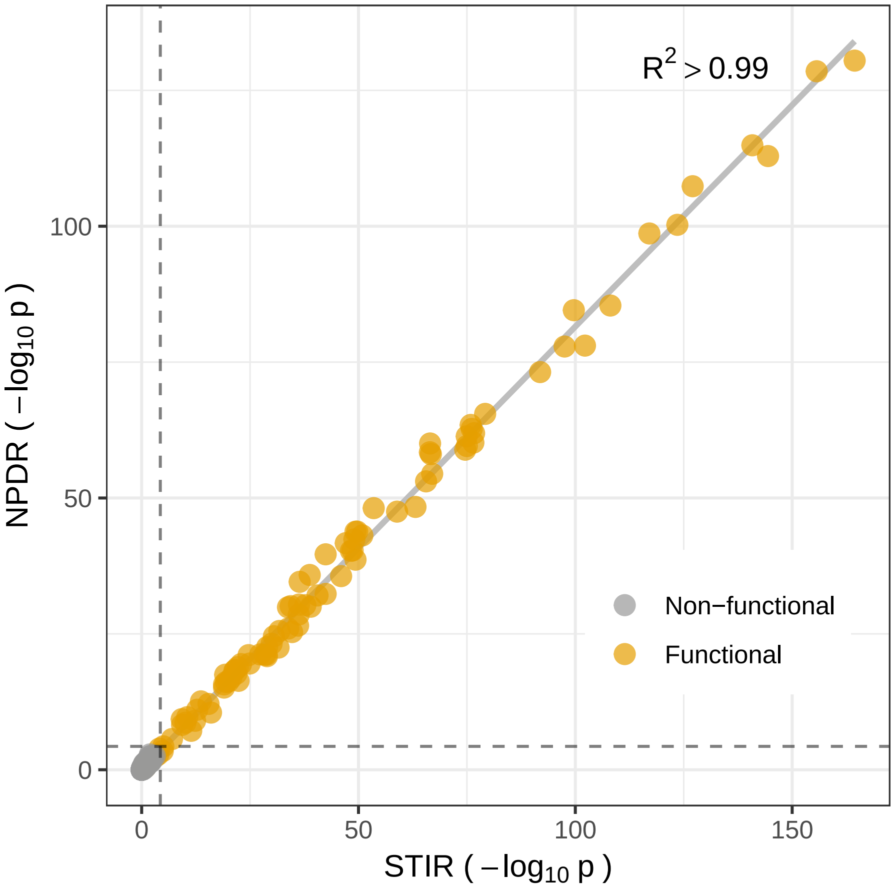

Figure S1: Similarity between NPDR and STIR for one simulation of $m = 200$ samples and $p = 1000$ attributes. In 100 replications, R^2 ranges from 0.9827 to 0.9994. STIR is based on a t-test of projected distances and NPDR is based on a logistic regression of projected distances.

References

[1] Trang T Le, Ryan J Urbanowicz, Jason H Moore, and Brett A McKinney. Statistical inference relief (stir) feature selection. *Bioinformatics*, 2018.

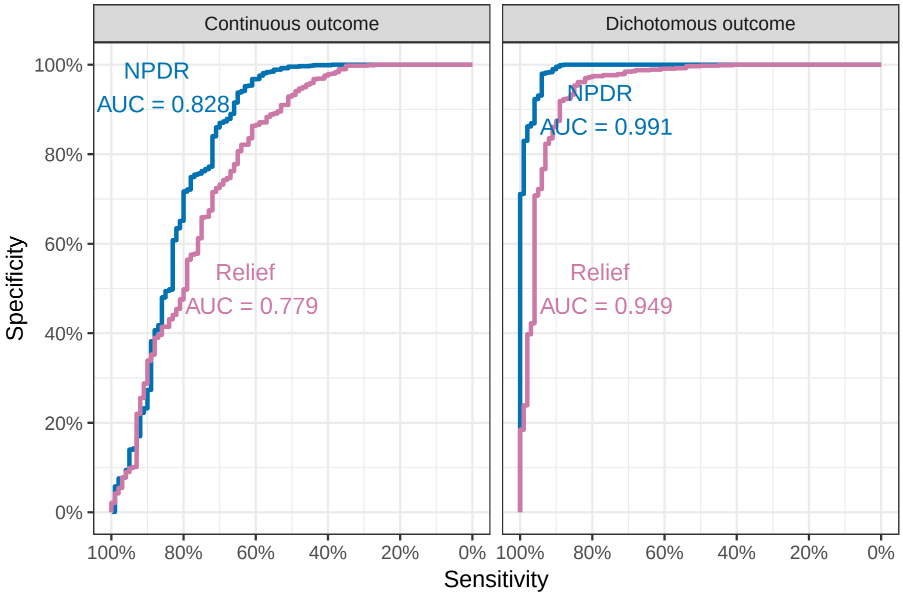

Figure S2: Receiver Operating Characteristics (ROC) curves for Relief-F and NPDR for simualted case-control data with interactions (left) and RRelief and NPDR for simulated continuous outcome data with main effects (right). Simulation uses $m = 200$ samples and $p = 1000$ attributes with 100 functional.

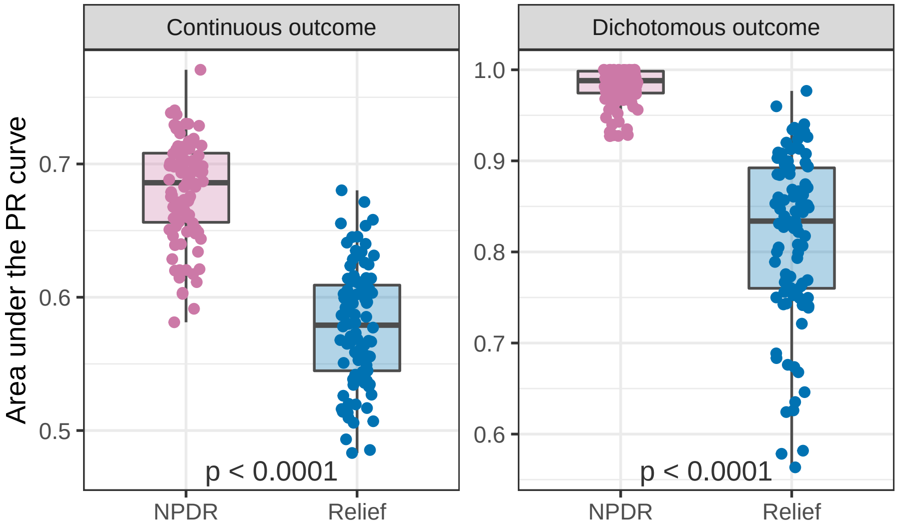

Figure S3: NPDR and Relief comparison of area under the PRC for 100 replicate simulations of case-control (left) and continuous (right) data. All simulations use $m = 200$ samples and $p = 1000$ attributes with 100 functional. NPDR yields significantly higher auPRC.

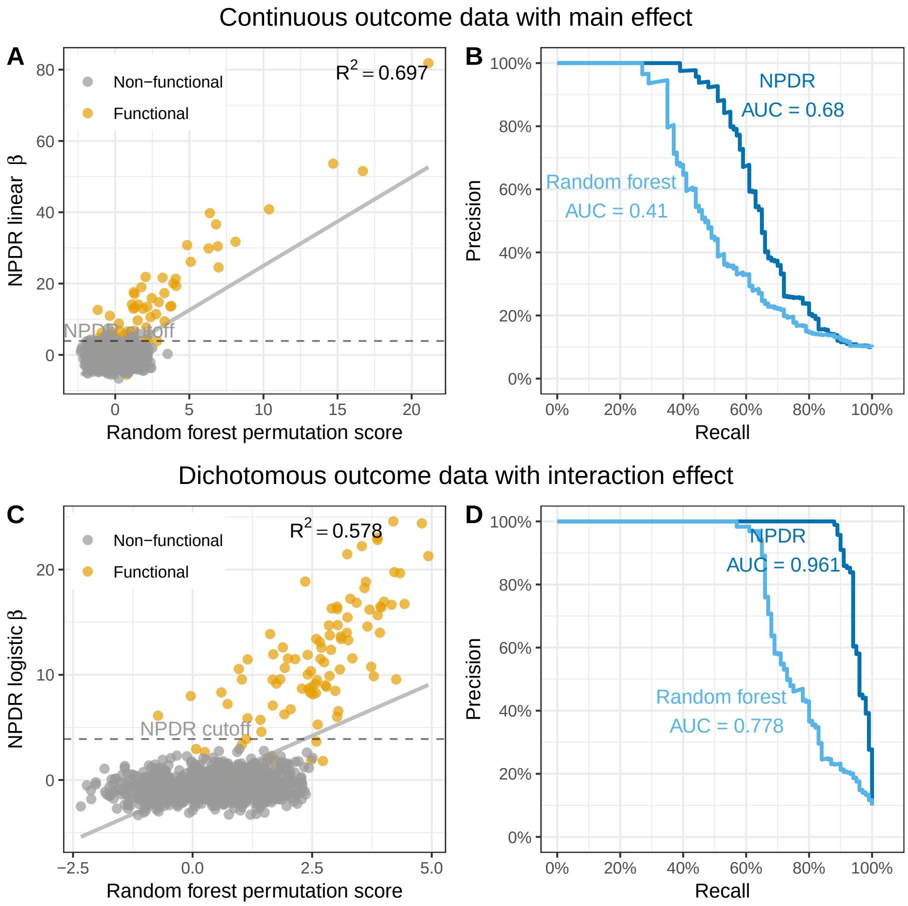

Figure S4: **Comparison of NPDR and random forest importance scores** for continuous outcome data with main effects (top row) and dichotomous outcome data with interaction effects (bottom row). Results for one replicate simulation ($m = 200$ samples and $p = 1,000$ attributes with 100 functional). For continuous outcome (A), importance scores computed by random forest permutation (percent increase in MSE) and NPDR standardized linear regression coefficient. For case-control outcome (C), scores computed by random forest permutation (mean decrease in accuracy) and NPDR standardized logistic regression coefficient. A regression line between the scores with R^2 is shown, and a 0.05 Bonferroni cutoff (dashed) is shown for NPDR (A and C). There is no statistical threshold for random forest, so area under the precision-recall curve (auPRC) is used to compare algorithm performance (B, D).

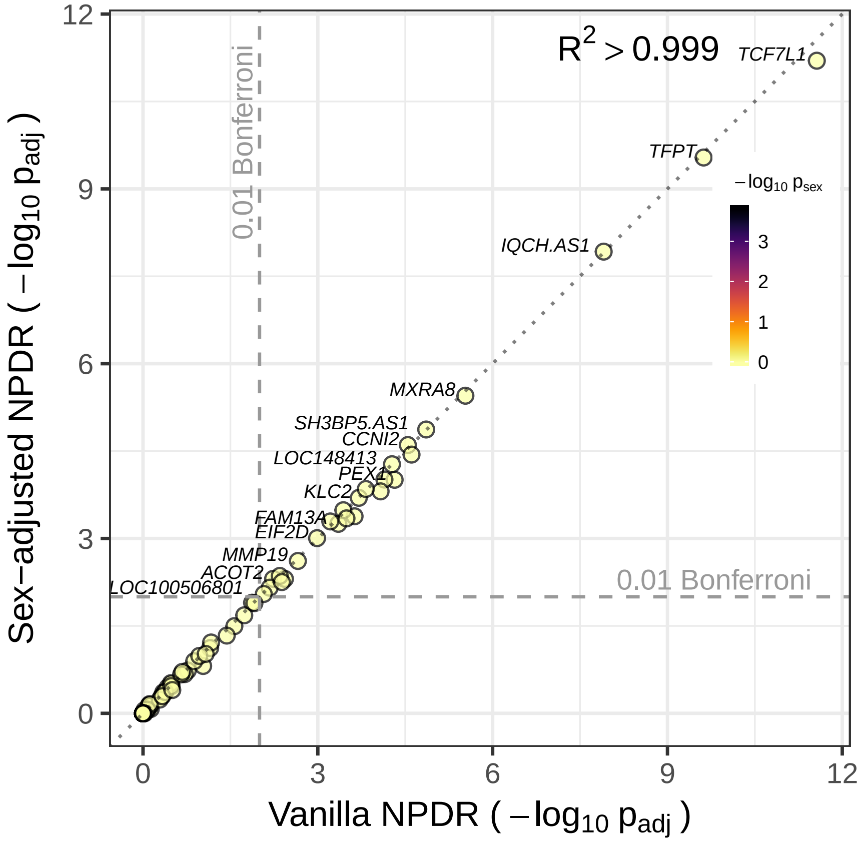

Figure S5: NPDR with and without sex adjustment for analysis of MDD-associated genes in Le et al.'s RNASeq dataset. Adjustment for the sex covariate has a negligible effect on the resulting P values for each important gene because of the balanced study design. Both methods yield consistent results with STIR from previous study (Fig. 4 of Ref. [1]), not shown.

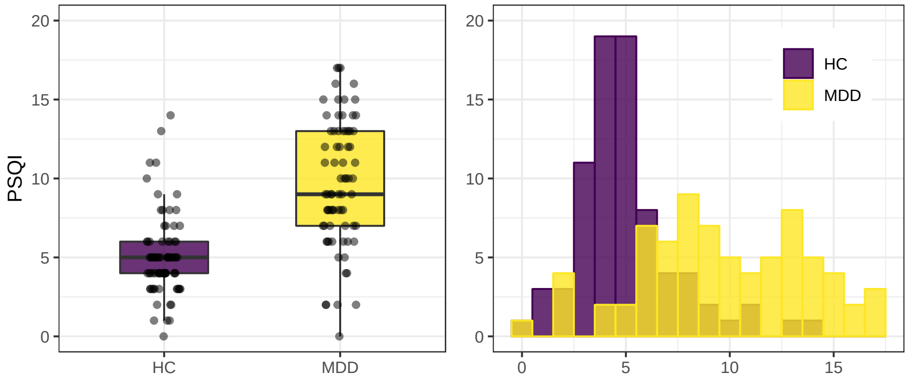

Figure S6: The distribution of the Pittsburgh Sleep Quality Index (PSQI) among individuals with and without MDD.