SUPPLEMENTARY MATERIAL

Supplement C: Comprehensive Review of Simulation Results Comparing Group-wise Association Tests

(doi: [10.1214/17-AOAS1028SUPPC](https://doi.org/10.1214/17-AOAS1028SUPPC); .pdf).

We provide a review of relevant results in previous research comparing group-wise association tests. Generalizing these results to our context implies that burden tests may have increased power to detect association at the block level when the direction of code-specific effects are similar. In contrast, when code-specific effects may differ in direction, SKAT has been shown to be an effective testing strategy [Wu et al. (2011)]. When provider-level clustering is present, the L-KM method controls the type I error and increases power compared to competing methods [Qi et al. (2015)].

Independent Data. Wu et al. (2011) compared the type I error and power of SKAT and burden tests for binary outcomes. In their simulation settings, genetic variants have a variety of frequencies: in the context of CPT code, this corresponds to data with both rare and common codes within a given block. Among the different types of burden tests considered, the counting-based burden test corresponds to our procedure in Section 2.2.1. The sum test by Basu and Pan (2011) also shares the same spirit with our burden test.

Both SKAT and burden tests are able to control the type I error at $\alpha = 0.05$ level. However, they can be conservative when the sample size is small. In terms of power, Wu et al. (2011) considered two different cases: (1) all the causal variants have positive effects and (2) 50% - 80% causal variants have positive effects and the rest are negative. For genetic applications the magnitude of the effect is assumed to be determined by the frequency of the variant: rarer variants have larger absolute effects (in either direction). When effects have different directions, SKAT has the highest power while the burden test suffers from low power. Not surprisingly, when all of the variants have effects that are in the same direction, the counting-based burden test has the highest power.

In practice with CPT codes, the choice between the burden test and SKAT for block-level inference should be guided by any available biomedical knowledge on the direction of code effects. Specifically, if within a block certain procedures are heavily used in one cohort, and some are heavily used in the other cohort, e.g., code substitution, then SKAT would be preferred.

Correlated Data. Qi et al. (2015) considered longitudinal outcome and compared the longitudinal kernel-machine regression (L-KM) method to other treatments of longitudinal outcome such as taking the average within cluster

and treating the average outcome as independent. The L-KM controls the type I error at numerous significance level, and has the highest power as it uses full information in the data.

References.

- BASU, S. and PAN, W. (2011). Comparison of statistical tests for disease association with rare variants. *Genetic Epidemiology* **35** 606-619.
- QI, Y., WEEKS, D. E., TIWARI, H. K., YI, N., ZHANG, K., GAO, G., LIN, W., X., L., W., C. and N., L. (2015). Rare-Variant Kernel Machine Test for Longitudinal Data from Population and Family Samples. *Hum Hered* **80** 126-138.
- WU, M. C., LEE, S., CAI, T., LI, Y., BOEHNKE, M. and LIN, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89** 82-93.