Author’s response to reviews

Title: The effects of electronic medical record phenotyping details on genetic association studies: HDL-C as a case study

Authors:

Logan Dumitrescu (logandumitrescu@gmail.com)
Robert Goodloe (robert.goodloe@gmail.com)
Eric Farber-Eger (eric.h.farber-eger@vanderbilt.edu)
Jonathan Boston (bostonou@gmail.com)
Dana C Crawford (crawford@chgr.mc.vanderbilt.edu)

Version: 2 Date: 1 December 2014

Author’s response to reviews:

Reviewer 1:

Major Compulsory Revisions

Overall this is a very interesting and necessary exploration of whether different clinical phenotyping algorithms used on EMR data provide the same overall results from association studies. The authors concluded that after they used 5 different algorithms along with a GRS to test for association with LDL-C, that the effect sizes were similar, suggesting that the algorithm used makes little difference. While these findings seem robust, I think the paper and analyses could be further strengthened by going beyond just the 7 index SNPs used to calculate the GRS. My question is whether there is actual correlation at the pvalue level for all genotyped snps across the 5 algorithms. Are the rank order of snp associations more or less the same regardless of algorithm used? That might address the question of whether the different algorithms impact discovery potential from GWAS studies of EMR data.

We thank the Reviewer for this interesting suggestion. We performed tests of association between all common variants on the Metabochip and each of the five HDL-C definitions as suggested. Unfortunately, the sample size is too small to perform reliable comparisons across the results for the discovery tests of association. None of the tests of association were significant after correction for multiple testing (as opposed to the associations between HDL-C and genetic risk score, which representreplications). We have added a description of this limitation of the study to the end of the Discussion section:

Another limitation of the present study is related to sample size and power. We present here tests of association between various HDL-C derived variables and an unweighted GRS. The unweighted GRS, by design, is calculated by the number of risk alleles at loci known to be significantly associated with HDL-C levels. Therefore, with only a few thousand samples, we were able to statistically replicate the expected association between the unweighted GRS and the various
HDL-C variables to further examine the genetic effect sizes estimated from these tests of associations. While the sample size of the present study was large enough for replicating known associations such as the loci represented in the unweighted GRS, the sample size is not large enough to perform discovery studies with the entire Metabochip dataset, even when limited to common variation (minor allele frequency >5%). Indeed, tests of association between the various HDL-C variables and common variants on the Metabochip failed to identify a statistically significant association after correction for multiple testing (data not shown). Furthermore, neither significance rankings nor genetic effect sizes could be reliably compared across HDL-C variables given the chance findings of non-significant tests of associations. Larger sample sizes are needed to make comprehensive comparisons of genetic effect sizes and significance rankings for EMR-derived phenotypes susceptible to algorithm decision logic and phenotyping details.

Minor Essential Revisions

Some minor spelling issues were found.

We thank the Reviewer for this comment. We have re-read the document and have corrected the spelling issues.