Editor Comments:

Please provide more details on the connection between the two types of responses (survival and logistic). For example, in the simulation study, after the event time and censoring time are generated, how did you convert it to a binary outcome so a logistic model can be adopted directly? Was the logistic model adjusted for age? It's not surprising to see the better performance of Cox model in simulation since it was the underlying true model. Will you still observe this pattern under mis-specification when the logistic model is the true one?

Current simulation study is limited. 30000 SNPs and 29,000 individuals are simulated each time but this is quite different from the empirical study (795850 SNPs and 49792 individuals). It is not very realistic to assume that the sample size and number of SNPs are of the same order in a GWAS. Please conduct a thorough simulation study combining a variety of sample sizes, number of snps, frequency of risk alleles, nominal levels (0.01 and 0.1 besides 0.05), etc. Also, why generating censoring time from gamma and a fixed cutoff? What is the data generating model?

>>> We have now scaled up the simulations, both in terms of number of SNPs and number of individuals, to match the empirical data. We have also simulated data from both a Cox model and a logistic model, and have used multiple p-value cutoffs. As before, our simulations included an empirically-derived distribution of allele frequencies. Across all these simulations, our finding that Cox regression has improved power still holds.

>>> Regarding censoring distribution, previous work has shown both theoretically and numerically that the inference procedure of the Cox model does not depend on the censoring distribution as long as it is non-informative. The Gamma distribution implemented in the paper is a parametric distribution that is non-informative and allows non-uniform censoring (Li 2018). The upper bound, i.e., fixed cutoff, for censoring time was used to represent the fact that empirical studies tend to have a maximum length of follow-up (e.g., the time at which the data are analyzed), referred to as administrative censoring. In the simulations from the logistic model, the censoring time was simulated from a uniform distribution (Hong et al. 2018).

Will the strategy of applying Cox after logistic regressions lead to biased results? Does it work well in simulation? If the order is switched, do you get similar results? And what if the two models are fitted not dependent on each other? What is the reason of modelling age as a cubic smoothing spline with 3 knots? It seems not done in simulation. If this is only modelled as a linear term, the logistic model might detect more associations.

>>> We have now applied the sequential strategy of logistic followed by Cox to our simulations, and have observed a similar improvement in sensitivity relative to logistic regression alone (Fig. 1). The two models were fitted independently of each other in our analyses, indicating that applying Cox regression on a subset of SNPs identified by logistic regression does not lead to biased results. We see no compelling reason to reverse the order of the sequential strategy,

since fitting the Cox model is much slower than fitting the logistic model. Indeed, the reason for using logistic regression first is to make the analysis time more tractable as the number of genotypes and phenotypes are scaled up.

>>> The spline gives the logistic model flexibility to fit the non-linear diagnosis rate vs. age (see examples in Figure 3), for which a linear term is inappropriate. We have now included splines in the analyses of the simulated data.

It might be a good idea to provide the QQ plot for genome-wide p-values. Besides, is it possible to provide details on how many associations are detected by the two models for the 50 phenotypes so we can better understand Figure S3? D and E of Figure S3 are confusing. The number of cases and controls are extremely unbalanced for all the phenotypes. Did this cause difficulty in convergence for logistic model? Is it still sensible to apply logistic model directly on the unbalanced data? Or maybe the influences can be assessed in simulation?

>>> We have added QQ-plots (Figure S2) to go with the Manhattan plots (Fig. 2). We have added the number of statistically significant associations detected by each method. We have also relabeled the axis titles of Figure S3D-E for clarity.

>>> Imbalance between cases and controls is a common feature of GWASes based on clinical data, because only a fraction of patients have any given phenotype. This type of study design has been validated in numerous studies, so studying the issue by simulations is outside the scope of the current manuscript. We only included phenotypes with ≥ 100 cases and SNPs with ≥ 1% minor allele frequency, and did not have any problems with convergence of the logistic model.

I agree with the reviewer that results on comparing odds ratio and Hazard ratio should not be over-interpreted.

>>> Please see our response to reviewer 2's comment below.

A typo on line 160.

>>> The misrendered character has been fixed.

Reviewer reports:

Reviewer 1: In the current study, the author compared Cox proportional hazard regression and logistic regression in the analysis of genotype-phenotype associations. Comparing to logistic regression which are more commonly used in the study of genotype-phenotype associations, Cox proportional hazards regression can account for times at which events occur. The results showed that Cox regression have greater power at equivalent Type I error. Therefore, it has advantages over logistic regression for longitudinal health-related data.

(1) Figure 1 looks confusion and misleading. What is the line represent? It does not look like a line of y=x. The labels for the number of simulations are also confusion and not clearly explained.

>>> We have revised Fig. 1 to show boxplots of the true positive rates for Cox and logistic regression.

(2) It is still easier for the readers to follow if the regression models that are used for the analysis is clearly presented in the paper, although the current paper did not propose new methods but evaluated established methods.

>>> We have clarified the regression models for both the simulated data and the empirical data.

(3) What happens if proportional hazard assumption does not met? Maybe the simulation study should consider situation that proportional hazard assumption does not met.

>>> We agree that in targeted studies, one should check the proportional hazards assumption and with evidence of non-proportionality, one should consider extended methods, including inclusion of parameters for time-dependent effects, changing to a different type of model, stratification on a prognostic factor exhibiting non-proportional hazards, separate modeling for different time periods, and weighted estimation and inference for the regression parameters (Wei and Schaubel 2008 and Schemper et al. 2009, among others).

>>> The ability to evaluate whether the effect varies by time, i.e., whether the hazard is non-proportional, is an advantage of Cox regression compared to logistic regression. However, the goal of the current simulations is to evaluate the performance of the Cox and logistic models when both are applicable, and additional simulations of non-proportional hazards are beyond the scope of the current paper. We will evaluate the performance of extended Cox models, as well as scalable methods to check the proportional hazards assumption, in future work.

Reviewer 2: In the manuscript 'Cox regression increases power to detect genotype-phenotype association in genomic studies using the electronic health record', Jacob Hughey et al compared the power of logistic regression and cox regression and recommended to use cox regression to increase power. If the time to event (time to an appearance of phenotype in electronic health record) is well defined, the cox regression will incorporate additional granular time information into test to increase power. The following a few points need clarification.

1. In simulation study, paired t-test was used to compare the true positive rate between logistic and cox regression. As the number of simulation can be easily increased or decrease, the test is over-powered. The CI and paired t-test could be excluded.

>>> We have removed the p-value, but have chosen to keep the confidence interval, as it indicates the degree of uncertainty given the number of simulations we actually ran.

2. Odds ratio from logistic regression and Hazard ratio from Cox regression are measuring different rate of developing disease with different assumptions. The magnitude of the numbers should not be compared.

>>> We agree that the hazard ratio and odds ratio are measuring different things, and we are no longer comparing their magnitudes. We have changed our description of the relationship between the two quantities from “nearly identical” to “strongly correlated”. We have also mentioned in the Discussion that the strong correlation between the odds ratio and hazard ratio is likely due to low event rates and a valid proportional hazards assumption. We note that previous work on genetic association studies compared odds ratios and hazard ratios as well (van der Net et al. 2008).

3. In application using electronic health records, it will be more informative giving detailed definition of event to time, overall survival type curve for a specific phenotype code.

>>> We have clarified Fig. 4 and its caption.