A genome-wide two-component mixture model expectation-maximisation algorithm for time to event data.

Ben Francis [1], Peng Yin[1], James Cook[1], Andrea Jorgensen [1], Jane Hutton [2], and Andrew Morris[1]

[1] University of Liverpool, United Kingdom.

[2] University of Warwick, United Kingdom.

Traditional survival analysis of time to event (TTE) data assumes that all individuals will experience the event of interest (EOI). In pharmacogenetics studies, there are patients who will not experience the therapeutic effect of a drug regardless of dosage or duration of prescription. Those who are unable to experience the EOI are deemed to be the part of the “cure fraction”.

Modelling TTE data consisting of those susceptible to the EOI and the cure fraction requires a “two-component” approach; enabling estimation of the effect of covariates on both susceptibility and the time to the occurrence of an EOI. One widely-used method incorporates an accelerated failure time model and expectation-maximisation algorithm but is too computationally intensive to be applied genome-wide. To circumvent this problem, we obtained survival and susceptibility residuals from a model including clinical covariates only, to then use as phenotypes in a linear regression model and a multivariate “reverse regression” analysis.

To assess the performance of these approaches, we performed detailed simulations incorporating a range of models of SNP effect on survival and susceptibility. Under a null model of no association of a SNP with survival or susceptibility, the type I error rates of all analytical approaches were maintained. Over the range of association models considered, the multivariate reverse regression approach was more powerful than linear regression for survival and susceptibility and no less powerful than the full two-component model. In conclusion, we have developed a novel “approximate” computationally efficient approach to enable genome-wide analysis of two-component TTE data.