The Alan Turing Institute

Forest-based **causal machine learning** methods identify drivers of **treatment effect heterogeneity** in trials well

A comparison of methods for detecting treatment effect heterogeneity in randomised controlled trials

Ellie Van Vogt[1,2], Karla Diaz-Ordaz[1,3]

[1]The Alan Turing Institute, [2]Imperial College London, [3]University College London

Introduction

RCTs typically aim to estimate the average treatment effect, however, there is increased interest in estimating the variation in treatment effects observed in RCTs. This is particularly o interest where the populations baseline characteristics have greater variation, a situation with greater generalisability of the results but potentially less homogeneity in the treatment effect, which might effect estimation methods used. We sought to estimate the conditional average treatment effect (CATE), that is, the treatment effect based on some baseline characteristics. We re-analysed the TRACT trial, comparing standard and liberal transfusion volume in children with severe uncomplicated anaemia in Uganda and Malawi. The primary outcome was 28-day mortality.

Methods

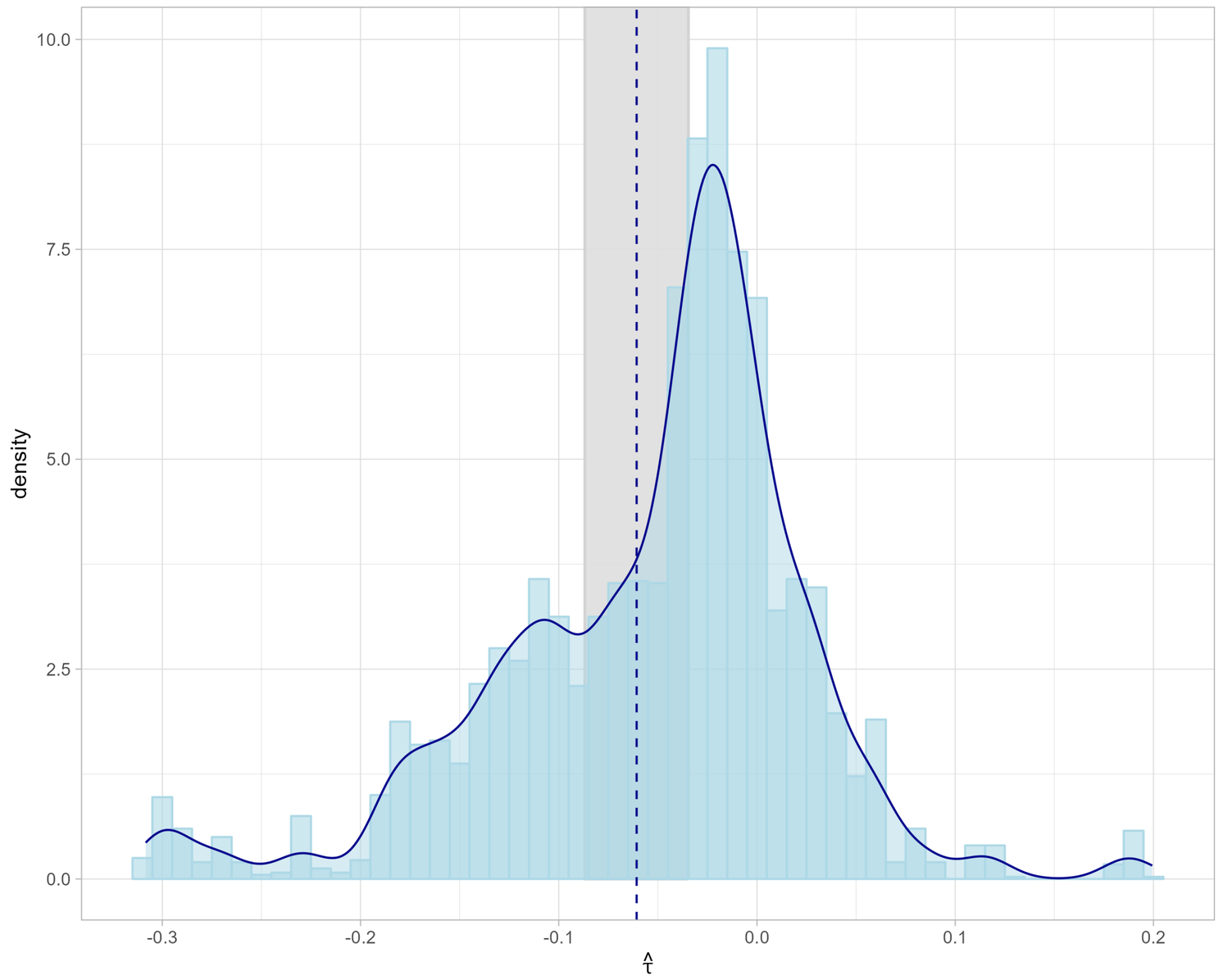
We compared four different models for the estimation of the CATE in the TRACT trial. The first models were meta-learning algorithms. We used the two step, T-learner approach with logistic regression models as base learners, with and without lasso penalty. The other models used were the causal forest (CF) and the Bayesian causal forest (BCF). Both of these approaches are causal machine learning (CML) methods. The CF and the BCF seek to estimate the CATE by using an ensemble of many trees to partition observations into groups with similar within-group treatment effects and large between-group differences in treatment effects.

There were 16 categorical variables highlighted by trial investigators as being potentially influential on treatment effect. An additional analysis was conducted where temperature indicators were exchanged for the continuous temperature measurement. Further, we did analysed a set of 42 variables of general interest to trial investigators. Training/test split was 70:30.

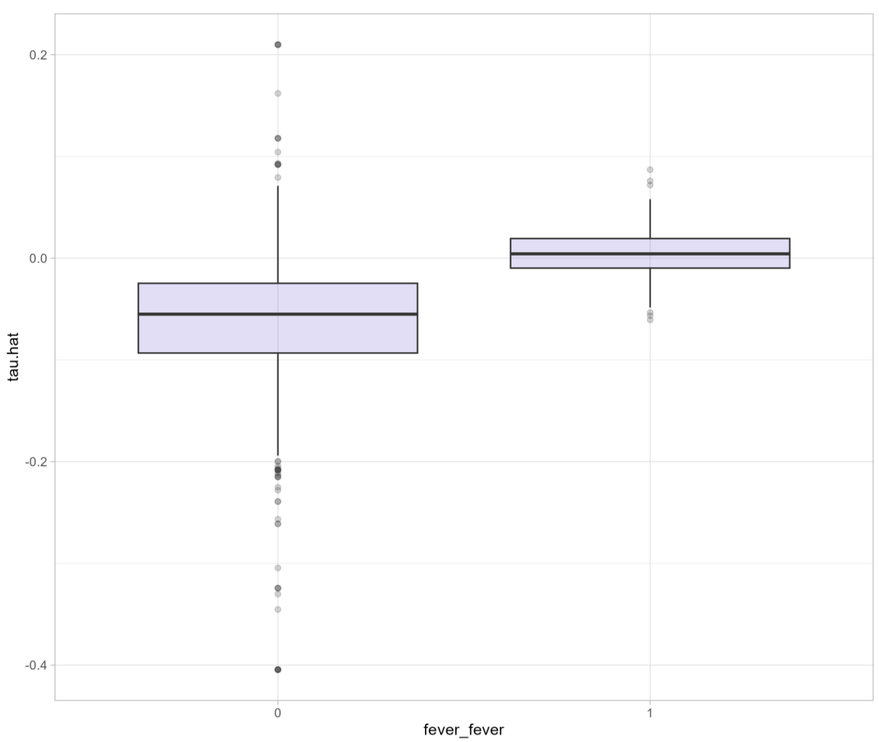
The primary outcome was rare (3.4% over total population). Therefore, we ran models both with and without upsampling in the training data. We also compared four different missing data handling approaches: complete case; mean imputation; random forest-based single imputation and inverse probability weighting (IPW).

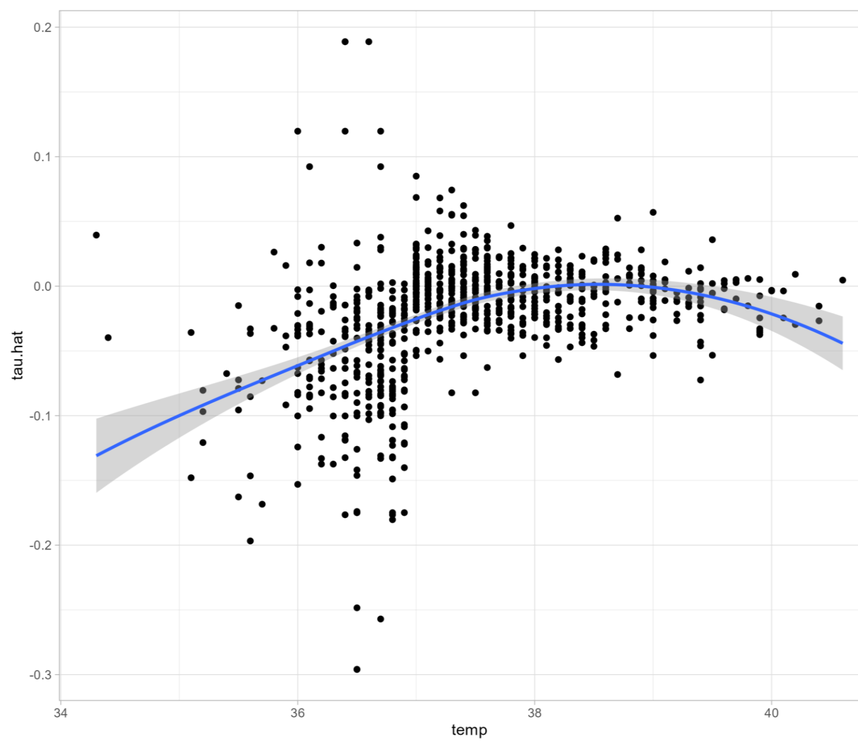
To understand which baseline variables were driving HTE in our population, we used variable importance measures. We compared standard variable importance measures with permutation variable importance methods. Permutation variable importance shuffles the observations of one dependent variable and then uses the magnitude of the change in the outcome as a measure of variable importance.

The CF package in R comes with an inbuilt function that acts as a generic, “omnibus” test for heterogeneity in a dataset once a model has been run. Analogous tests were devised for the BCF and T-learners using the best linear predictor framework.



Top:out of bag CATE estimates as generated by the causal forests with 16 categorical baseline variables and IPW missing data handling. The CATE is risk difference in 28-day mortality. Left: CATE in test data participants without (0) and with (1) at baseline, using 16 baseline variables dataset and IPW missing data handling. Right: test data CATE estimates in dataset with 15 categorical baseline variables and temperature and missing data handling with IPW





Results

The forest-based models found variation in treatment effect with the presence or absence of fever. Children without a fever had a negative treatment effect, meaning they had a decreased risk of mortality by receiving a liberal transfusion volume over the standard transfusion volume (left). The same was observed in the dataset including temperature (right). This pattern was observed for all model setups.

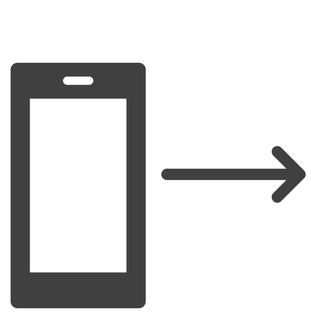
Upsampling training datasets improved the accuracy of CATE estimation and HTE detection however upsampling caused difficulties in BLP estimation as randomisation had been broken. In particular, T-learner approaches were less able to accurate model CATE, particularly without the upsampling of the outcome. (See QR code).

Some analyses were not able to be completed, such as IPW missing data handling for BCF as weights are not currently compatible with available models.

Discussion

In this project we found that forest-based methods were superior to T-learners with logistic base-learners for treatment effect estimation in our data. Upsampling rare outcomes was beneficial to improve detection of drivers of heterogeneity however, further investigation is required on the assumption violations associated with breaking randomisation and having propensity scores potentially insufficiently bounded away from 0 and 1.

Scan for more on this project



UKRI MRC Clinical Trials Unit

UCL

Imperial College London