Ideas for simulation setup for OOS Estimation:

* Different distributions depending on the study
  + Use a few of the MDD variables and shift them
  + Options
    - All distributions are the same
    - Each study sample has a slightly different severity of depression min and max
    - Two studies have heavier and older patients
    - The studies are fully differentiable in terms of their age distribution
* Different CATEs depending on the study
  + Same setup as before – study doesn’t matter, study matters a little bit, study matters a lot

Simulation data

* Setup used before from Tan and Kunzel papers
* MDD data
  + Use the causal forest to define a CATE model within each study and apply that to everyone and assume that’s the truth?
    - Then fit a CF with study indicator
  + Define the CATE based on covariates and study and get each person’s outcome based on that
    - Then fit a CF with study indicator
    - For the extra people, use the same CATE form but replace the study coefficient with some random noise?? Or add in something to represent that they are from observational data – new unobserved confounder?
  + **Make up own setup but using MDD data**
    - **First fit outcome model – just a linear regression – with main effect and interaction effect terms**
      * **Use that to inform the m and tau functions**
    - **Maybe I can use the EHR data to get a cate function for the test data? Or just use the linear regression but include a random error or a new unobserved confounding relationship**

When each approach will likely work well:

* Random approach
* Membership model approach: when there is a clear distributional difference between studies
* Default GRF approach

Evaluation

* MSE
* Confidence interval coverage
* Confidence interval length

Questions

* How do we define the CATE for the individuals OOS? Same function, or just add in some random error due to not knowing study?