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 more severely depressed 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**
    - **Ask liz about how to decide final model**

Function from Vivli for scenario 1a:

* Y = 10.7 – 8.0\*Study132 – 12.7\*StudyT304 – 10.3\*StudyT315 – 8.5\*Vorti – 0.02\*Age – 0.87\*MADRS\_bln -0.15\*female + 0.15\*Study132:MADRS + 0.44\*StudyT304:MADRS\_BASE + 0.38\*StudyT315:MADRS + 0.07:Vorti:Age + 0.20\*Vorti:MADRS + 3.04\*Study132:Vorti + 3.01\*StudyT304:Vorti + 0.59\*StudyT315:Vorti
  + So m = 10.7 – 8.0\*Study132 – 12.7\*StudyT304 – 10.3\*StudyT315 – 0.02\*Age – 0.87\*MADRS\_bln -0.15\*female + 0.15\*Study132:MADRS + 0.44\*StudyT304:MADRS\_BASE + 0.38\*StudyT315:MADRS
  + Tau = – 8.5 + 0.07\*Age + 0.20\*MADRS + 3.04\*Study132 + 3.01\*StudyT304 + 0.59\*StudyT315 (NOT DIVIDED BY 2…)

Function from Vivli for scenario 1b (nonlinear):

Function from Vivli for scenario 2 (study-specific functions):

Step by step of code for simulating data and trying methods:

1. Simulate training data using gen\_mdd() function
   1. Vary up the distribution of the covariates (same, madrs varies, madrs and age vary, age completely distinguishable across studies)
   2. Vary up the tau function (linear or non-linear, both derived from real MDD data)
   3. **Check on weight distribution**
   4. **Is it okay that I’m just doing + W\*tau instead of the (2W-1)/2\*tau?**
2. Simulate test data
   1. Just do a grid of values but for each get an actual tau function based on the scenario but removing study
   2. Vary up the test tau function (random error or an unobserved confounder)
3. Fit model to training data using causal forest with pooling with trial indicator (?)
   1. Get training MSE just to have (?)
4. Predict on testing data according to each imputation method and create confidence intervals for each test individual
   1. Completely random
   2. Study membership model
   3. Within-forest default
   4. Within-forest random sampling
5. Calculate metrics of accuracy
   1. MSE between estimated and true CATE
   2. Confidence interval coverage
   3. Confidence interval length

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?