Section 1:

The next day, Dr. Thorin introduces you to Dr. Bombur, an attending pediatrician at MGH. Dr. Bombur has studied that low birth weight (lbw) might have a negative effect on a baby’s chances of one-year survival. He is considering starting an initiative that helps pregnant women take better care of their own health so that their babies are born at a healthy birth weight. Before embarking on this initiative, Dr. Bombur wants to make sure that lbw does indeed have an effect on infants’ health, particularly their chances of survival. He asks you:

“Does low birth weight really decrease a baby’s chance of survival beyond his/her first birthday?”

For this task, you will use the lbw data. This is a publicly available dataset of linked birth and infant death data collected and maintained by the Center for Disease Control. Each line represents an infant, and has data about the infant’s date of birth and death if he/she died within a year. It also has information about the mother, her education, race, smoking habits, alcohol consumption and other as well as some information about the father.

Download the data here:
http://people.csail.mit.edu/dsontag/courses/mlhc_summer18/day2/singletons.csv
http://people.csail.mit.edu/dsontag/courses/mlhc_summer18/day2/twins.csv

Note that in this part of the question (Q1x), we are looking at birthweight under 2700 grams as the treatment = 1 and 0 otherwise, and the outcome is mortality.

Q1a: Dr. Bombur mentioned that his previous research assistant suggested studying the effect of low birth weight in the population of twins. Why did the previous RA think that the twin population is particularly well suited for the task? Which causality assumptions does the twin population likely satisfy? Use the lbw data to estimate the ATE of low birth weight on one year mortality. To do so, you need to use twins where only one of them is below the threshold. Does this ATE generalize to the whole population including singletons? Meaning: can we assume that the ATE of lbw in the singletons population is roughly the same? Hint: compute the mortality rates among the twins and the singletons population.

Q2: Dr. Bombur realized he can refine his question a bit. He explains that his initiative would directly target smoking cessation among pregnant women, since it has been shown that babies born to women who smoke during their pregnancy are more likely to have a lbw. If a mother’s smoking habits truly do cause babies to be born at a lower weight, Dr. Bombur’s initiative would be useful. For the remainder of this lab, use the singleton population for your analyses.
Do a covariate adjustment (using linear regression) to estimate the effect of mother’s smoking habits on the baby’s birthweight (now using the continuous measure, not the binarization). Clearly state what your outcome is and which variables you include in the regression. Give an example of a feature that you should not include in your model. Finally, report the weight of the treatment. What should Dr. Bombu conclude?

Q3: An alternative to covariate adjustment is propensity score weighting. Compute the ATE after reweighing the distribution using propensity scores. Comment on the distribution of propensity scores using plots or summary statistics as you see fit. Compute and report the ATE.

After your meeting with Dr. Bombur, you thank all the physicians for their time and decide to stop by the Hospital cafeteria for some seed cake. It’s been a long day.

Section 2:

Choose one of these two papers for your group to read and discuss:

1) *Personalized Diabetes Management Using Electronic Medical Records*
   Dimitris Bertsimas, Nathan Kallus, Alexander M. Weinstein, and Ying Daisy Zhuo
   Diabetes Care, 2016
   [http://care.diabetesjournals.org/content/early/2016/12/01/dc16-0826.full-text.pdf](http://care.diabetesjournals.org/content/early/2016/12/01/dc16-0826.full-text.pdf)

2) *Medical Homes and Cost and Utilization Among High-Risk Patients*
   Susannah Higgins; Ravi Chawla; Christine Colombo; Richard Snyder; and Somesh Nigam
   American Journal of Managed Care, 2014

How does the causal inference problem that needs to be solved in the paper relate to what you learned in the lecture? Does it attempt to estimate the average treatment effect (ATE) or the individual treatment effect (ITE)? How does it use machine learning for this (e.g., through covariate adjustment, propensity score, or a variant of these)? What are potential weaknesses in the authors’ analyses? How do the results inform health care policy?