Getting started hints - for students

This section provides a few brief hints for the student in how to begin thinking about analysing the data.

Project 1 - PreECMO

First extract the variables that you will be using (i.e. remove all Day1ECMO variables from the dataset). Since we don't want to overfit our models, randomly splitting the data into training/validation (and possibly test) datasets will help us avoid this. Only split once and then save the separate datasets so that the same ones can be used for all models. An alternative to this is cross validation.

Since this is a binary classification problem with ECMO_survival as an outcome, initial examination of continuous variables as to their usefulness for prediction can be done with both visualisations and simple hypothesis tests. For categorical variables, simple tabulations and tests will be sufficient.

Be aware that there is missing data. You may choose to leave out observations with missing data but that may leave you with very few observations to work with. You may choose to remove variables instead and perform simple imputation on remaining variables with a small amount of missing data.

Many classification methods you have met are only for continuous variables, e.g. LDA, QDA, k-nearest neighbours, so be aware of this when applying them to your dataset.

Reading material
Hastie, T., Tibshirani, R. and Friedman, J. (2017) The Elements of Statistical Learning (2nd edition). Springer.

Carolyn S Calfee, Kevin L Delucchi, Pratik Sinha, Michael A Matthay, Jonathan Hackett, Manu Shankar-Hari, Cliona McDowell, John G Laffey, Cecilia M O'Kane, Daniel F McAuley (2018) Acute respiratory distress syndrome subphenotypes and differential response to simvastatin: secondary analysis of a randomised controlled trial. The Lancet Respiratory Medicine 6(9), pages 681-698.

Pratik Sinha, Kevin L. Delucchi, B. Taylor Thompson, Daniel F. McAuley, Michael A. Matthay, Carolyn S. Calfee (2018) Latent class analysis of ARDS subphenotypes: a secondary analysis of the statins for acutely injured lungs from sepsis (SAILS) study. Intensive Care Medicine 44(11), pages 1859-1869.

Project 2 - Day1ECMO

First extract the variables that you will be using (i.e. remove all PreECMO variables from the dataset). Since we don't want to overfit our models, randomly splitting the data into training/validation (and possibly test) datasets will help us avoid this. Only split once and then save the separate datasets so that the same ones can be used for all models. An alternative to this is cross validation.

Since this is a binary classification problem with ECMO_survival as an outcome, initial examination of continuous variables as to their usefulness for prediction can be done with both visualisations and simple hypothesis tests. For categorical variables, simple tabulations and tests will be sufficient.

Be aware that there is missing data. You may choose to leave out observations with missing data but that may leave you with very few observations to work with. You may choose to remove variables instead and perform simple imputation on remaining variables with a small amount of missing data.

Many classification methods you have met are only for continuous variables, e.g. LDA, QDA, k-nearest neighbours, so be aware of this when applying them to your dataset.

Reading material

Hastie, T., Tibshirani, R. and Friedman, J. (2017) The Elements of Statistical Learning (2nd edition). Springer.

Carolyn S Calfee, Kevin L Delucchi, Pratik Sinha, Michael A Matthay, Jonathan Hackett, Manu Shankar-Hari, Cliona McDowell, John G Laffey, Cecilia M O'Kane, Daniel F McAuley (2018) Acute respiratory distress syndrome subphenotypes and differential response to simvastatin: secondary analysis of a randomised controlled trial. The Lancet Respiratory Medicine 6(9), pages 681-698.

Pratik Sinha, Kevin L. Delucchi, B. Taylor Thompson, Daniel F. McAuley, Michael A. Matthay, Carolyn S. Calfee (2018) Latent class analysis of ARDS subphenotypes: a secondary analysis of the statins for acutely injured lungs from sepsis (SAILS) study. Intensive Care Medicine 44(11), pages 1859-1869.