**Copied from nathan's notebook**

Another thing to take into account is that although it is normally expected that ID values should not have influence over the training of a machine learning model, entries in this case can originate from the same patient, adding bias to the data. This would eventuate data leaks between the test and training data sets if patient ID is not taken into account when splitting.

In saying this, a possible solution to this would be to perform splits while ensuring that entries from the same patient ID are kept grouped together. For example, all entries from patients 1, 2, 3, 4, 5 can be set aside for training and entries from patients 6, 7, 8, 9, 10 can be set aside for testing (given that the number of entries within the split data sets are proportioned)