## Initial HES QC (not yet estimated):

The purpose of this step is to establish whether the correct positive patients have been pulled from the HES database and whether patients are correctly linked across tables.

For the positive patients, the QC will involve using the existing Sheffield database and checking that linked patients are matched terms:

1. Year of birth
2. Gender
3. Whether or not they had a right heart catheter (OPERTN field) at Sheffield (PROCODE field).

For all patients (positive and negative):

1. Extract the unique set of ICD codes (using both the primary and secondary fields; DIAG\_01 and DIAG\_{02,..}) and provide patient count for each. Join this with the reference table for the ICD codes and flag those that do not match a code in the reference table.
2. Extract the unique set of procedure codes (OPERTN\_\*) and provide patient count for each. Join this with the reference table for the procedure codes and flag those that do not match a code in the reference table.
3. Extract the unique set of main specialties (MAINSPEF) codes and provide patient count for each. Join this with the reference table for the MAINSPEF codes and flag those that do not match a code in the reference table.
4. Extract the unique set of hospitals (PROCODE) codes and provide patient count for each. Join this with the reference table for the PROCODE codes and flag those that do not match a code in the reference table.

Following these four steps the codes that do not have a match in the reference table (erroneous codes) should be assessed taking into account any special characters that may be errors and also the patient count of the erroneous code.

## ETL (2 weeks; confirmation required once size of data is known):

Briefly, the ETL will comprise taking the event-level HES data across three tables (APC, OP and AE) and creating variables which will capture events between a lookback (earliest) date and an index (most recent) date. These patient history variables will be primarily made up of ICD-10 codes, procedure (OCPS) codes and specialty visits. Patient counts from the positive cohort and negative cohort will be used to select variables according to their trigger rate (how often they are present in each cohort; this step will involve a team decision. Additionally, these variables can be binned in a temporal fashion (e.g. between 0 to 18 months before index date). The temporal binning of these variables will be driven by the descriptive stats of the patient history across the positive and negative cohorts.

A subset of variables such as age, ethnicity, mental health category and socio-economic status information will only be extracted at the index date.

For the detailed ETL plan please refer to:

****

## QC and Descriptive Stats (1 week):

Final QC of output from ETL.

Univariate and bivariate stats.

Correlation analysis.

Cap extreme values.

## Patient Finding (2 weeks):

Two cohorts will be available at this point – the positive iPAH cohort and the negative cohort. The negative cohort will comprise a true negative group who have been assessed at Sheffield (Xns) and a collection (Xnn) of nationally treated true negatives and nationally treated false negatives. In this stage of the project we would like to identify these false negatives and groups these with the true positives in Xps.

Take Sheffield defined true positives (Xsp) and true negatives cohort (Xsn; defined by the Sheffield data) and split each of these into two partitions.

Using the first partition of each cohort (Xsp1 and Xsn1), tune the SMOTE algorithm to model each cohort.

Sample two new cohorts from the SMOTE model to create Xsp1’ and Xsn1’.

Build a classifier to discriminate between Xsp1’ and Xsn1’.

Apply this model to the second partition of the cohorts (Xsp2 and Xsn2) to produce a PR curve. Select a threshold from this curve so that the FP rate is conservative.

Apply this model to the national-level negative cohort (Xnn) to select potential positive patients with a conservative FP rate.

## Variable creation (1 week):

Creation of key sequence or interaction variables (co-occurrence of events).

Creation of date difference variables.

## Modelling Phase 1 (3 weeks):

Design sampling structure guided by patient counts – either cross-validation with hold out or cross-validation only.

Design positive to negative ratio for a representative ratio.

Logistic regression (LR) will be used as the benchmark model.

LR with flag variables only

LR with flag and sequential.

LR with flag and sequential and date differences.

Random forest (RF) with coarse grid search.

RF with frequency and data differences and sequential.

Model evaluation using PR curves

Readily available variable importance computed using the model coefficients (e.g. gini coefficient for random forests or model coefficients of LR) as well the relative risk scores.

## Modelling Phase 2 (3 weeks):

Exact tasks to be decided based on insight from data and preliminary model.

Options could include:

Gradient Boosting Trees (xgboost).

Ensemble approach including LR, xgboost and RF.

Two stage modelling.

Clustering to group patients into more homogenous groups.

Further feature engineering.

Scope to be limited to what can be achieved in 15 days.

Enhanced patient finding.

Outputs:

Model evaluation using PR curves

Overall rank of variable importance.

Model interpretation in the form of relative risk scores.