ETL plan.

1. Prepare code to establish the index date. CODE PREPARED
   1. For PAH positive cohort, this is the date that a patient visits the specialty centre which is TCD by the PROCODE and MAINSPEF. Code should be prepared to input the specialties and providers via a list. If a valid index date is not found then use the criteria for the negative cohort.
   2. For PAH negative, the index date should be the latest date of the set of relevant ICD codes or MAINSPEF visit. Patients who do not have a valid index date should be excluded. Code should be flexible so that in can take these in as a list. *May require a rule to assess if the index date is earlier. For e.g. if a negative patient is late in their patient journey this might be sub-optimal.*
2. Prepare the code to establish the lookback date (earliest date of patient history). CODE PREPARED
   1. PAH positive – earliest date of a list of given ICD codes or visit to set of MAINSPEF. If these events are not observed it should be the date of the earliest entry in any table (APC or A&E or OP). No positive patients should be excluded at this point.
   2. For PAH negative, the index date should be the earliest date of the set of relevant ICD codes or MAINSPEF visit. Code should be flexible so that in can take these in as a list. Negative patients who do not have a valid lookback date should be excluded.
3. Compute descriptive stats on the length of lookback for the positive and negative cohort, separately. CODE PREPARED
   1. Include the histograms and the mean, sd, and percentiles at 0, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95 and 1.
4. Provide attrition numbers for the negatives patients. CODE PREPARED
   1. Number of patients who were excluded due to invalid index date.
   2. Number of patients who were excluded due to invalid lookback date.
5. Create table of patients and their row entries that fall within the relevant index and lookback dates. CODE PREPARED
6. Create counts (patient level) of variables for diag codes in the positive cohort and a random sample of the negative cohort (assuming that the negative cohort is very large). These counts should be provided across all three tables as well as an aggregated count. Select all diag codes that occur greater than x% of the time. Percentage to be determined after inspecting the table of counts. This should be performed for both the primary and secondary diag codes and computed across 4 levels of aggregation. CODE PREPARED  
   The sample OP table has 11 secondary diags and the APC table has 19.
7. Create count of variables for procedure codes in the positive cohort and a random sample of the negative cohort. These counts should be provided across all three tables as well as an aggregated count. Select all procedure codes that occur greater than x% of the time. Percentage to be determined after inspecting the table of counts. CODE PREPARED
8. Create count of variables for specialty visits in the positive cohort and a random sample of the negative cohort. These counts should be provided across all three tables as well as an aggregated count. Select all specialty visits that occur greater than x% of the time. Percentage to be determined after inspecting the table of counts. CODE PREPARED
   1. List of MAINSPEF numbers on index date
   2. List of MAINSPEF numbers on visit prior to index date
   3. Number of days between our assigned index date and the true index date derived from the Sheffield dataset. Anyone who has more than X days, flag it!

**OUTPUT – which diag/spec/procedure variables should be included in the final model; the index and lookback dates.**

1. Counts of the events (procedures, diag and specialty visits). Configure these using the following temporal ranges (TBC): CODE PREPARED
   1. 0 to 6 months from index.
   2. 6 to 12 months from index.
   3. 12 to 24 months.
   4. 24 month to lookback date.
2. Date of first exposure for all diag (primary and secondary), procedures and specialty visits. CODE PREPARED
   1. For e.g. use as the number of days since index.
   2. For e.g. create sequence variables (continuous form).
3. If trim\_level=3 and a code has length 2, should we pad it with zeros, leave it out, or accept it? NEEDS TO BE DISCUSSED
4. Determine how to code (and in some cases whether to include) the multi-level categorical variables: NEEDS TO BE DISCUSSED
   1. ADMIMETH – simple cat, one hot encode
   2. ETHNOS – simple cat, one hot encode, 99 is missing, should we remove?
   3. HRGNHS (pull for ETL but don’t include in the model?) - should we include this? There are 430 different values just in the mock data.
   4. HRGNHSVN (pull for ETL but don’t include in the model?)
   5. SUSCOREHRG - 874 unique values just in the mock
   6. MENTCAT – simple cat, 8 levels, 9 is unknown, include 9?
   7. RURURB\_IND – simple cat, - 9 levels
5. Calculate age as age at index date. For example, when an index date occurs in the APC table then ADMIAGE can be used. CODE PREPARED
   1. Produce descriptive stats for positive and negative cohorts.
6. Episode duration variable – this could be the total number of hospital days over the entire patient history. CODE PREPARED
   1. Have three variables – one that is the count of bed days from the APC table, another that is the count of episodes in the OP table and a final variable that counts the number of A&E visits.
7. Extract the each individual SES score at index data using as well as the overall rank. CODE PREPARED
8. Extract rural/urban index (RURURB\_IND) at index date. CODE PREPARED

**OUTPUT – FLAT FILE FOR THE POSITIVE AND NEGATIVE COHORTS.**