**RT-HaND\_C Strategy Summary for maintenance and growth of the database: Post-EPIC Cohort Data Curation**

This document lists a brief overview of the strategy to build prospective patients into the HNC clinical dataset. The prospective cohort comprises of patients seen by the HNC oncology team after 4/10/23, or those seen before but who did not complete their radiotherapy course until after 4/10/23. For full inclusion/exclusion criteria please see the full clinical dataset build SOP document.

**Step 1. Prospective patient identification**

The HN Masterlist is reviewed monthly and new patients updated since the last review are added to the ‘Prospective patient’ spreadsheet. This spreadsheet retains as much information as possible from that on the HN Masterlist to avoid data loss. This includes Hospital ID number, date of birth, name, date seen in oncology clinic and some clinical details including treatment plan documented on the HN Masterlist. Duplicate patients are then checked for and removed.

* This hospital ID list is then be checked on the Guy’s Data Warehouse (GDW) (FROM [Reporting].[rep].[VwPatientLegacyIdMappingDim) using SQL code to see if patients have legacy hospital IDs.
* Patients with legacy hospital IDs should then be checked against the retrospective cohort – and if present deleted from the prospective cohort list to avoid duplication.
* This prospective cohort list can then have clinical data populated from the GDW (Step 2) and imaging/radiotherapy data (see relevant SOP document).

**Step 2. Population of data for prospective cohort patients.**

The key source of data will be data tables on the (GDW), which is currently the data warehouse being used to store data from Epic. There is currently a limited amount of data we require on the GDW from Epic. The location of available data on the GDW, if known, is shown for each data category below (‘FROM…’).

***Current data identified:***

Alternative Medical Record Numbers (MRN)

* Legacy HospitalID (FROM [Reporting].[rep].[VwPatientLegacyIdMappingDim])
* NHS number (FROM [Reporting].[rep].[VwPatientDim])

Demographics

* Full demographics data (date of birth, sex, ethnicity, post code, address) (FROM [Reporting].[rep].[VwPatientDim])
* This can then be used to generate English indices of deprivation data for each patient’s post code (from <https://imd-by-postcode.opendatacommunities.org/imd/2019>)

Co-morbidities

* CogStack was used for the retrospective cohort
* Co-morbidity data is available from the GDW but will need extraction and cleaning into same format used for retrospective before useable (FROM [Reporting].[rep].[VwOutpatientDiagnosisFact])

Disease

* Some disease data including major site, laterality (if documented), histology (if documented) and disease staging (FROM [Reporting].[rep].[VwCancerStagingFact])

Treatment

* Overview (FROM [Reporting].[rep].[VwCancerTreatmentFact]) – we will need to ratify that this data is currently in a format we need
* MDT data (leg.somm\_VwSTG\_CancerMdmEventFact)

Outcomes:

* Dead/alive, date of death (FROM [Reporting].[rep].[VwPatientDim])

Other data categories may be available on the current list of GDW tables, but not yet available. Outstanding data tables will need to be requested from the GDW team. Key missing areas include:

* Chemotherapy treatment data
* Surgical procedure data
* Radiotherapy treatment data
* Other outcome data (e.g. last follow-up)
* Toxicity data (from H&N assessment form v5 and PROMs forms)
* Bloods data

Radiotherapy dose data will likely require extraction for each patient from Aria (Eclipse).

**Step 3. Amalgamation of retrospective and prospective clinical data.**

Once relevant data has been extracted for prospective patients, this can be saved as CSV files on the Radiotherapy Head&Neck Subject Area Mart (SAM) on the Electronic Data Warehouse (EDW), home to retrospective dataset. In such a way, data can therefore be amalgamated in relevant categories to the retrospective cohort.