**RT-HaND\_C SOP work instructions for maintenance and growth of the database: Pre-EPIC Cohort**

This document lists areas of missing data for the HNC clinical dataset. Items are listed in priority order (high to low).

As of 4/10/23 (the date to which the retrospective cohort data is up-to-date currently), around 50% of patients in the cohort were still alive. Future data for these patients will also therefore require curation – primarily further treatment and outcome (disease and toxicity) data categories. This data will likely need to be extracted from the Guy’s Data Warehouse, the data warehouse containing Epic data, when it becomes available there.

Additionally, a completeness assessment should be completed every 6 months (see main SOP document for completeness assessment methodology).

**Priorities for retrospective cohort data curation:**

1. **Disease outcomes and updated follow-up dates**

Currently the most prominent missing area, as reflected in the table below. Priority is therefore given curating data for these data categories. Where possible, some data for these categories has been curated by abstraction from other data points (e.g. patients with neck dissection post-RT identified as having a salvage neck dissection). However the likelihood is that the majority of this data will require manual curation.

|  |  |  |  |
| --- | --- | --- | --- |
| **Cause of death** | 2682 | 213 | 7.4% |
| **Cause of death other** | 2737 | 158 | 5.5% |
| **CogStack CR and radical RT** | 2561 | 334 | 11.5% |
| **3m RT response** | 1153 | 1742 | 60.2% |
| **3m imaging modality** | 1991 | 904 | 31.2% |
| **3m imaging date** | 1621 | 1274 | 44.0% |
| **6m RT response** | 1317 | 1578 | 54.5% |
| **6m imaging modality** | 1926 | 969 | 33.5% |
| **6m imaging date** | 1797 | 1098 | 37.9% |
| **Biopsy post RT** | 2895 | 0 | 0.0% |
| **Bx date** | 2895 | 0 | 0.0% |
| **Salvage neck dissection** | 2204 | 691 | 23.9% |
| **ND date** | 2216 | 679 | 23.5% |
| **Failure** | 1831 | 1064 | 36.8% |
| **Failure date** | 2113 | 782 | 27.0% |
| **Failure site summary** | 2103 | 792 | 27.4% |
| **Primary failure site details** | 2239 | 656 | 22.7% |
| **Primary recurrence date** | 2219 | 676 | 23.4% |
| **Primary recurrence Rx** | 2257 | 638 | 22.0% |
| **Primary recurrence intent** | 2257 | 638 | 22.0% |
| **Nodal or locoregional failure site** | 2226 | 669 | 23.1% |
| **Nodal or locoregional non primary recurrence date** | 2220 | 675 | 23.3% |
| **Nodal recurrence Rx** | 2257 | 638 | 22.0% |
| **Nodal recurrence intent** | 2257 | 638 | 22.0% |
| **Metastatic site** | 2221 | 674 | 23.3% |
| **Met recurrence date** | 2184 | 711 | 24.6% |
| **Met recurrence Rx** | 2188 | 707 | 24.4% |
| **Met recurrence intent** | 2224 | 671 | 23.2% |
| **Recurrence after RT** | 1991 | 904 | 31.2% |
| **Time to failure after RT** | 1991 | 904 | 31.2% |
| **In field recurrence high dose** | 2135 | 760 | 26.3% |
| **In field recurrence ppx dose** | 2135 | 760 | 26.3% |

Follow-up data (death date and last follow-up date) can also be updated for retrospective cohort patients when data available on GDW. Date that outcomes are updated to must be recorded to ensure dataset consistency.

1. **Disease features**

Disease feature data is a further prominent area with large amounts of missing data, as reflected in the table below. CogStack’s performance was not currently adequate for population of a number of these concepts such as ‘lymphovascular invasion’. The likelihood is that the majority of this data will require manual curation by review of legacy EHR documents/clinical notes/histopathology reports.

|  |  |  |  |
| --- | --- | --- | --- |
| **Disease subsite** | 1705 | 1190 | 41.1% |
| **Disease laterality Mosaiq** | 1618 | 1277 | 44.1% |
| **pT stage Mosaiq TNM7** | 2626 | 269 | 9.3% |
| **pN stage Mosaiq TNM7** | 2624 | 271 | 9.4% |
| **pDisease Stage TNM7** | 2726 | 169 | 5.8% |
| **cT stage Mosaiq TNM8** | 2616 | 279 | 9.6% |
| **cN stage Mosaiq TNM8** | 2616 | 279 | 9.6% |
| **cM stage Mosaiq TNM8** | 2617 | 278 | 9.6% |
| **cDisease Stage TNM8** | 2625 | 270 | 9.3% |
| **pT stage Mosaiq TNM8** | 2724 | 171 | 5.9% |
| **pN stage Mosaiq TNM8** | 2724 | 171 | 5.9% |
| **pDisease Stage TNM8** | 2726 | 169 | 5.8% |
| **Histopath Date** | 2895 | 0 | 0.0% |
| **Tumour Grade Mosaiq** | 2615 | 280 | 9.7% |
| **HPV p16** | 2895 | 0 | 0.0% |
| **HPV ISH** | 2895 | 0 | 0.0% |
| **GT LVI** | 2726 | 169 | 5.8% |
| **Perineural invasion Combined GT and CogStack data\*** | 0 | 2895 | 100.0% |
| **Perineural invasion present CogStack\*** | 0 | 2895 | 100.0% |
| **GT PNI** | 2726 | 169 | 5.8% |
| **Surgical margin** | 2726 | 169 | 5.8% |
| **Depth of invasion** | 2740 | 155 | 5.4% |
| **GT ECS** | 2726 | 169 | 5.8% |

\* Perineural invasion (PNI) could be manually curated at same time as other data categories to improve accuracy, as current data relied on CogStack, which may be less accurate.

Also required:

* Checking diagnosis dates (e.g. gap between surgery and Dx date etc) – looking at certain differences
* Filling in missing Decision to Treat dates

1. **Future treatments**

Patients in the retrospective cohort may go on to receive additional treatment:

* Surgery
* Radiotherapy
* Chemotherapy

Once available from the GDW, this extra data needs to be incorporated into the retrospective cohort data.

1. **Toxicity data**

* Toxicity (H&N assessment v5 and PROMS) forms from Epic also need to be incorporated once available on the GDW for retrospective cohort patients.
* All data categories on HN forms need to be added to dataset, but not all on there currently.
* We need to apply abutting data ranges to capture all forms and prevent any missing data. Currently applied date rules applied for data capture:
* Baseline:
  + Forms done before or after RT start date
* On-treat forms:
  + Week 1 – days 2-7
  + Week 2 – days 8-14
  + Week 3 – days 15-21
  + Week 4 – days 22-28
  + Week 5 – days 29-35
  + Week 6 – days 36-42
* Post-treatment forms:
  + 1 week-post – 1 to 7 days post-RT completion
  + 2 weeks-post – 8 to 14 days post-RT completion
  + 6 weeks-post – 35 to 49 days post-RT completion
  + 3 months-post – 70 to 110 days post-RT completion
  + 6 months-post – 152 to 213 days post-RT completion
  + 9 months-post – 243 to 304 days post-RT completion
  + 1 year-post – 305 to 456 days post-RT completion
  + 2 years-post – 638 to 821 days post-RT completion
  + 3 years-post – 1003 to 1186 days post-RT completion
  + 4 years-post – 1338 to 1582 days post-RT completion
  + 5 years-post – 1703 to 1947 days post-RT completion
  + 5 years and more-post – all forms from 1703 days and after post-RT completion

1. **Other diagnostic test data**

* Audiogram data
* Bloods – other date points (e.g. 1 year post-RT), thyroid function tests
* EDTA GFR

1. **Missing radiotherapy data**

* Completion of clinical trial data by manual curation or cross-referencing with other sources.
* Completion of planning MRI data by manual curation or cross-referencing with other sources.
* Complete interruption to treatment data by manual curation.