**Radiotherapy Head and Neck Clinical Database (RT-HaND\_C)**

Last updated: 15/11/2024

**Introduction**

**Importance of Real-World Data in Oncology**

To develop robust Deep Learning (DL) toxicity prediction models we require large and accurate clinical and radiotherapy (RT) imaging datasets (1). Relevant clinical data includes demographic, disease, treatment, and outcome data, whilst RT imaging data necessitates consistent and accurate organ-at-risk (OAR) delineation. At our centre, over 2000 Head and Neck Cancer (HNC) patients have been treated with Intensity Modulated RT (IMRT) since its introduction in 2011. If accurately curated, data from these patients can provide the large datasets required to train and evaluate effective DL toxicity prediction models for ultimate incorporation into the HNC RT workflow. Furthermore, a comprehensive and accurate clinical and imaging dataset of all HNC oncology patients can provide a wealth of data to help answer many other important unanswered research questions.

The US Food and Drug Administration (FDA) defines Real-World Data (RWD) as “data on patient health status and/or delivery of health care routinely collected from multiple sources outside typical clinical research settings” (2) such as Electronic Health Records (EHRs). RWD can be analysed to generate Real-World Evidence (RWE). There has been a growing interest in increasing the application of RWE within the field of oncology (3), with Cancer Research UK’s latest data research strategy highlighting the RWD’s capability to improve patient outcomes (4). This drive is underpinned by several factors including the limitations of clinical trials, the principle form of gold standard practice-determining research within oncology. Clinical trials provide vital insight into interventions’ abilities to improve patient outcomes but lack inclusivity, with potentially limited generalisability to the whole patient population, particularly with an increasingly co-morbid, older patient population, often excluded from studies. Furthermore, trials are increasingly expensive to perform and may require follow-up for many years before providing answers to research questions (5). RWE derived from RWD therefore has a role in conjunction with clinical trial research. Leading examples of large oncology RWD sources being used to generate RWE include the USA’s Surveillance, Epidemiology and End Results (SEER) Programme, which consists of 18 registries collecting cancer data for 48% of the US population (6), and the UK’s National Cancer Registration and Analysis Service (NCRAS), collecting data from multiple sources for all patients living in England diagnosed with cancer (7).

**RWD in Oncology: Context and Challenges**

Advancements in the integration of IT systems within UK healthcare, including the widespread adoption of EHR systems across the NHS (8), have increased the availability and accessibility of RWD sources, to a scale previously impossible to achieve. Whilst increasing use of EHRs gives rise to the potential for rapid and automatic extraction of data, curation of RWD still poses many challenges. RWD are collected during routine clinical practice, rather than for research purposes and can therefore be messy, inconsistent, incomplete and subject to random and non-random recording errors and biases (9). There are ongoing initiatives to improve consistency and quality of oncology data at the EHR data recording level, such as the American Society’s Clinical Oncology’s Minimal Common Oncology Data Elements (mCODE) initiative (Figure 1), which may facilitate consistency of terminology and data recording across different centres to aid robust RWE generation (10).

A diagram of a patient

Description automatically generated

Figure 1. mCODETM Version 4, detailing key RWD oncology data elements for routine collection within clinical practice (11).

RWD fragmentation also poses a significant challenge, with cancer patient data often held across multiple EHR systems, providing challenges to develop interoperable data pipelines. For example, when undertaking an oncology outpatient clinic at our centre a clinician routinely required data from at least five different EHR systems (Cancer Information System (Mosaiq), Electronic Patient Record (EPR), imaging system (PACS), outpatient appointment system (Patient Information Management System (PIMS)), radiotherapy planning system (ARIA)) for each patient review prior to the Trust’s move to a new EHR (Epic). Epic has amalgamated the old Cancer Information System, outpatient appointment system and EPR, reducing the number of systems a Clinical Oncologist routinely has to access during a patient consultation to three.

A further issue arises from the fact that a large proportion of important data within the EHR exists in an unstructured format (12), for example as text within a histopathology report. Structured data can be extracted from EHR systems using scripts or employing data platforms such as HealthCatalyst, which can facilitate rapid and automatic data mining from a data warehouse containing patient data across many different EHR systems (13). The gold standard for extracting unstructured data for real-world datasets is through human abstraction (manual curation), however this may not be practical for large datasets due to time and therefore resources required to carry out such an operation at scale (3). As well as time-consuming, manual curation of unstructured data can also be error-prone (14). Artificial Intelligence (AI) tools using Natural Language Processing (NLP), such as CogStack, have been developed to curate data rapidly and accurately (15), but there has been limited application of NLP technologies to mine big cancer data (16). CogStack’s ability to mine unstructured data for HNC patients has been evaluated previously, with clear strengths and limitations to consider when applying it to the curation of a clinical dataset.

Given the unique challenges underpinning RWD curation and subsequent use for RWE, robust validation of data quality is required. The growing use of RWD for RWE generation, and the necessity to ensure maintenance of data standards, has been reflected by publication of RWD data quality guidance by several regulatory bodies including the UK’s National Institute for Clinical Excellence (NICE) (17), the FDA (18) and the European Medicines Agency (19). Although the terminology described varies, these guidelines outline consistent necessary data quality dimensions that RWD requires including dataset validity, completeness, uniqueness, accuracy, consistency and timeliness. Castellanos et al. from Flatiron Health, a commercial company that provides RWD curation services to healthcare organisations, categorised the necessary validation categories RWD must achieve into “relevance” and “reliability” criteria (3) (Figure 2). Relevance of RWD can be subdivided into availability of critical variables, as well as sufficiency and representativeness of population. Reliability can be subdivided into accuracy, completeness, provenance, and timeliness. In keeping with regulatory body guidance, these data quality dimensions must be formally assessed and met by RWD before subsequent use to generate reliable RWE with valid conclusions.

A diagram of data quality

Description automatically generated

Figure 2. RWD quality dimensions (figure adapted from Castellanos et al.). (3)

**HNC Oncology Clinical Dataset Development (RT-HaND\_C)**

As a busy NHS oncology centre, we have available a vast amount of key clinical and imaging (diagnostic and radiotherapy) data. The majority of existing RWD sources (whether available or not on request to other researchers), including SEER and NCRAS, do not currently contain the annotated imaging data required for DL RT toxicity modelling research linked with clinical data. Plans for the introduction at our centre of the new unifying EHR system on 5th October 2023, Epic, to replace legacy EHR systems highlighted the necessity to build a clinical dataset for our patients treated prior to Epic go-live whilst legacy EHR data remains easily accessible. Once Epic data warehousing solutions are fully integrated, this dataset can be linked with data pipelines to both update with prospectively collected data for existing patients (such as follow-up or subsequent treatment data) and add future HNC oncology patients (Figure 3).

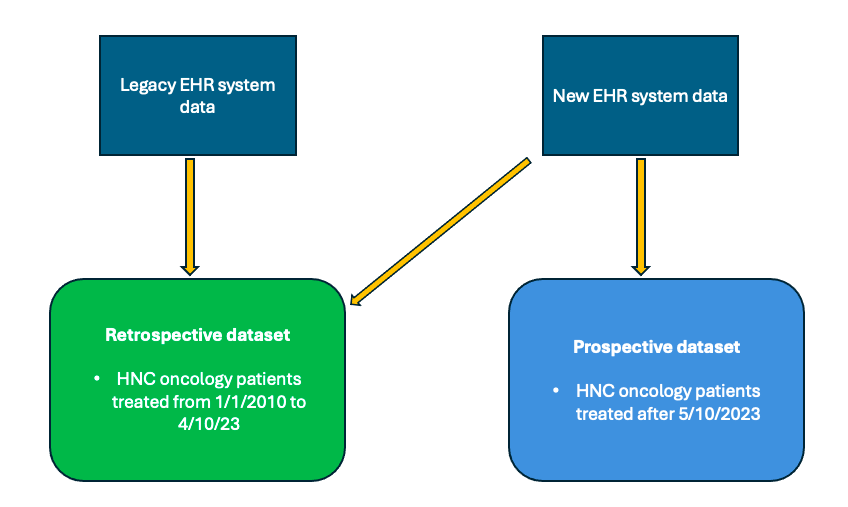


Figure 3. Retrospective and prospective dataset construction schema.

Following RWD validation guidance set out above, rigorous validation and verification checks were performed on our clinical dataset throughout and after dataset construction. Our team’s clinical expertise was used alongside database construction skills including the use of SQL for EHR structured data queries for data curation. Following clinical dataset (RT-HaND\_C) construction, clinical data can be linked to each patient’s relevant diagnostic and radiotherapy imaging/treatment data (RT-HaND\_I) to facilitate further multi-modal research projects including DL toxicity modelling. This is performed using the Extensible Neuroimaging Archive Toolkit (XNAT), an “extensible open-source imaging informatics software platform dedicated to imaging-based research” (20) (Figure 4) and the process for construction of this database is described elsewhere (Document 011 Transfer of Imaging Data SOP and Document 013 Transfer of Radiotherapy Data SOP). XNAT allows secure and rapid access to multi-modal (imaging and radiotherapy) data. The focus of this Standard Operating Procedure (SOP) report is the curation of a complete and accurate retrospective clinical dataset to 5th Oct 2023. However work is ongoing to develop an automatic clinical data curation pipeline from EHR systems for a prospective clinical dataset of all future HNC oncology patients.

A diagram of a patient population

Description automatically generated

Figure 4. Workflow for HNC RT patient clinical and imaging datasets.

**Aims of this work**

1. Build a retrospective clinical dataset of demographic, disease, treatment and outcome data for all HNC patients treated during the “IMRT era”, from (1/1/2010) to Epic go-live date (5/10/2023) (RT-HaND\_C)
2. Populate this database to completion using EHR queries (structured data points) and CogStack output for validated concepts (unstructured data points).
3. Validate the clinical dataset as per established data quality dimensions.
4. ‘Virtual linkage’ of the clinical database (on Electronic Data Warehouse (EDW)) with patients’ corresponding diagnostic imaging and radiotherapy planning and treatment imaging (RTHaND\_I on XNAT).
5. Set up a researcher SOP to provide ongoing validation and updating of the existing retrospective dataset.
6. Set up a prospective database tool to automatically update retrospective clinical dataset patient outcome data and collect data for all new HNC patients after Epic go-live.

**Methods**

**Population Definition: Inclusion/Exclusion Criteria**

Inclusion and exclusion criteria were set for the HNC oncology retrospective cohort to represent the range of HNC patients attending the HNC oncology clinic and therefore to minimise population bias.

* *Inclusion criteria*
  + Patients seen by the HN oncology team from 1/1/2010 to 4/10/2023.
  + All patients with head and neck primary cancers (nasopharynx, hypopharynx, oropharynx, oral cavity, paranasal sinus/nasal cavity, larynx, HN carcinoma unknown primary (CUP), salivary gland, lacrimal gland, multiple HN primaries).
* *Exclusion criteria*
  + Thyroid, parathyroid, skin cancer primaries (including auditory canal), paraganglioma tumours, lymphoma.
  + Patients that received all oncological treatment elsewhere.
  + Cases completing primary RT course after 4/10/23 (these patients will be included in the prospective cohort dataset).
  + Patients with initial treatment pre-2010 unless recurrence or new primary more than 5 years later.

The HNC Masterlist (detailed below) contained all *episodes of care* for patients seen by the HN oncology team since 2009. This list was reviewed and validated after the Epic go-live data to finalise the cohort based on the above inclusion and exclusion criteria. During this process, any duplicate patients were removed and incorrect Medical Record Numbers (MRNs) were identified by running a CogStack output for the whole population and reviewing those where zero clinical documents were identified. Any incorrect MRNs were corrected as facilitated using the HNC Masterlist’s date of birth and name data. This list of MRNs for our target population was cross-checked with several other sources to ensure no patients were missed. For example, the MRN list was compared to data of all HNC RT patient treatment records obtained from ARIA (over 2500 patients). Our final cohort MRN list contained only 6 patients (a single week accounted for 3 patients) not present on the HNC Masterlist, identified from RT treatment records (4 patients) and PD-L1 data (2 patients). In total, a cohort of 2895 patients was finalised as the complete population of interest using the above process.

**Data sources**

Before data curation was undertaken for our patient population, required data categories were identified. Demographics, co-morbidities, disease, treatment and outcome data categories were all identified for data population. A strategy for each data category was then established from the sources set out below (Figure 5). Wherever possible, automatic data curation methods at scale were favoured.

A diagram of a clinical dataset

Description automatically generated

Figure 5. Data sources used for the retrospective clinical dataset.

1. *HNC Masterlist*

As discussed above, the HNC Masterlist was the principal data source used to establish our cohort. This is a spreadsheet maintained by the HNC Clinical Oncology Consultants recording all patients reviewed by the HNC oncology team (inpatient and outpatient) since 2010. This document had been rigorously updated twice weekly since its creation in 2009 and also contained relevant information beyond patient details (MRN, name, date of birth) including intended initial treatment, RT fractionation, and RT start/finish dates. From this document not only was our patient cohort established but also basic clinical information (for example tumour site, staging, initial planned treatment, intended radiotherapy fractionation, start date and end date) was obtained to corroborate with other sources such as structured data extraction during the data curation process.

1. *Structured Data Extraction from EHRs*

Extraction of data from several EHR systems was possible for automatic and rapid curation of clinical data for structured data categories. This process was facilitated by HealthCatalyst, a platform utilised at our centre to hold vast amounts of clinical data tables from legacy EHR systems on a single EDW. Available data tables on the EDW were reviewed to identify those containing relevant data required for the population within our clinical dataset. Using the EDW, data could be extracted and data tables combined using Structured Query Language (SQL) queries on Microsoft SQL Server Management Studio (SSMS). All EDW data tables of interest were updated on 4/10/23 and therefore up-to-date from our retrospective data cohort’s required final assessment time point. The EHR systems containing data extracted in this method were:

* Mosaiq, used within our centre as the primary source for oncology documentation for cancer patients also containing chemotherapy and radiotherapy treatment information.
* EPR, our centre’s primary EHR hosting patient results and documents.
* PIMS, our centre’s primary EHR patient appointment data.

When required, for example for data cleaning purposes, data tables obtained from SQL queries on EDW were saved as CSV files and cleaned, before being reloaded to the EDW using an “Extract, Transform, Load” (ETL) data curation process (21). Transformed data was loaded as CSV files to the EDW using HealthCatalyst’s Subject Area Mart Designer tool (SAM Designer), allowing these loaded data tables to then be queried alongside other EDW data tables. This functionality was also used to load some legacy EHR data tables not currently ingested onto EDW, which were obtained directly from specific EHR systems and then uploaded to the EDW via SAM Designer. For example, HNC patient NCI CTC v.4.0 and v.5.0 and RTOG toxicity assessment form data from Mosaiq, collected prospectively at the point of care, was not originally present on the EDW’s available Mosaiq data tables, but was provided by our centre’s Mosaiq IT team in CSV form and then uploaded to the EDW. A similar process was used for Mosaiq chemotherapy data and Mosaiq RT request forms. The HealthCatalyst team were also able to update the EDW with any missing specific data tables on request if table location within the specific legacy EHR was provided.

1. *Data Extraction using CogStack*

Previous work by our team evaluated CogStack’s NLP tool to establish data categories that CogStack could reliably populate clinical data for. CogStack was run for our patient population across all relevant clinical documents within EHR systems for validated concepts. This provided an output of the number of times a concept was picked up within the specified searched EHR system documents for each patient. Using a concept-specific thresholding approach, this output was then converted to a binary output of “0” (concept not present) or “1” (concept present) for each concept for each patient depending on whether the threshold concept presence was met. This output was then loaded to the EDW as a CSV file (via SAM Designer) and could then be integrated with other relevant data tables using SQL queries when building our dataset.

CogStack’s output did not directly provide temporality of a concept’s presence (for example date of a procedure or treatment) nor detail beyond the presence or absence of a concept within the patient’s entire EHR. The limitations of CogStack’s output therefore dictated what data categories it could be appropriately used for within the context of the dataset build. The full list of SNOMED CT concepts run for the patient cohort can be seen in Appendix 1.

1. *Manually Curated Data*

For some data categories it was not possible to curate data using any of the above approaches, but some data was available from previously manually curated datasets, which were then integrated into the retrospective clinical dataset. For example, staging within the EHR was recorded as clinical TNM7 staging, but pathological TNM7 staging as well as clinical and pathological TNM8 staging were also of interest but not recorded in a structured format. Our department had undertaken a number of retrospective research projects, where disease and outcome data had been curated. These retrospective datasets comprised 172 oral cavity cancer patients (22), 69 larynx cancer patients (unpublished data) and 524 oropharynx cancer patients (23). This valuable source of previously manually curated data was therefore built into our dataset. This “Ground Truth” datasets (referred to in this report as “pre-existing GT datasets”) were less complete for certain data categories than those populated with one of the above data curation sources due to the specific nature of these retrospective research projects. Despite this weakness, as an available source of accurate data it would have been wasteful not to utilise such a resource, as it could be integrated with data from other methods to improve data completeness and accuracy for that population subset.

A further example of a manually curated dataset used for the dataset build was for PD-L1 status/score data, which was otherwise neither stored in structured form on the patient’s record nor represented adequately by current SNOMED CT concepts for mining using CogStack. However, our centre’s histopathology team had kept a complete record of HNC PD-L1 testing and kindly shared this data, which was also loaded to the EDW via SAM designer.

**Data curation**

Data curation source was dependent on the availability of data for each data category. Recognised NHS data definitions (24) were followed as closely as possible to ensure data consistency and standardisation. To maximise data reliability, any inaccuracies or inconsistencies within any of the data categories below that were observed during the data curation process were corrected from data sources files through to finalised data tables on SAM Designer. Data curation strategy for each data category is described below. A data dictionary (see Appendix xx– Data Dictionary) was also developed describing all data categories recorded in terms of description, primary data source, data type and coding format. Due to the nature of the different data points missing data points varies across different headings and is not static due to ongoing updates carried out either by the HN Database team or individual researchers.

*Demographics*

EPR contained all core patient demographic data and this data was therefore obtained from the relevant data table using SSMS on the EDW for all patient cohort MRNs. The following data was obtained in this way:

* NHS number
* Surname
* Forename
* Date of Birth
* Gender
* Ethnicity/Race

Post code and address were extracted from the EDW’s Mosaiq demographics data table using an SQL query containing all patient cohort MRNs. Mosaiq was used for this data as it was observed to most likely contain the patient address at the time of cancer diagnosis as opposed to during treatment or follow up (for example moving to a nursing home at the end of life) compared to other EHR systems. Address at cancer diagnosis was felt to be more relevant to future research needs. The first address line was also recorded within our dataset to ensure patients listed as living in care homes/hospices could be easily identified if required and original address checked easily for future work where required.

Further demographic data was extracted from patient assessment tables from Mosaiq using SSMS on the EDW for all patient cohort MRNs:

* Height (cm)
* First recorded weight (kg) and date recorded
* First recorded ECOG Performance Status and date recorded
* Alcohol intake status (current/previous/never/no) and date recorded
* Alcohol intake amount (</= 14u per week, >14u and </= 21u per week, >21u per week) and date recorded
* Smoking status (current/previous/never/no) and date recorded
* Smoking amount (< or = 10 Pack Years, >10 Pack Years) and date recorded

Postcode data was used to obtain English indices of deprivation 2019 data. Cohort postcodes were entered into the UK Government lookup tool (25) to generate all required data. This data was then downloaded as a CSV file and loaded as a data table to the EDW using SAM Designer. Post code deprivation data could then be linked via SSMS to each patient by linking on postcode. The data categories acquired were:

* Postcode data: Lower Layer Super Output Area (LSOA) code (a geographic unit for reporting of small-area statistics)
* Postcode data: LSOA Name
* Postcode data: Index of Multiple Deprivation Rank
* Postcode data: Index of Multiple Deprivation Decile
* Postcode data: Income Rank
* Postcode data: Income Decile
* Postcode data: Income Score
* Postcode data: Employment Rank
* Postcode data: Employment Decile
* Postcode data: Employment Score
* Postcode data: Education and Skills Rank
* Postcode data: Education and Skills Decile
* Postcode data: Health and Disability Rank
* Postcode data: Health and Disability Decile
* Postcode data: Crime Rank
* Postcode data: Crime Decile
* Postcode data: Barriers to Housing and Services Rank
* Postcode data: Barriers to Housing and Services Decile
* Postcode data: Living Environment Rank
* Postcode data: Living Environment Decile
* Postcode data: Income Deprivation Affecting Children Index (IDACI) Rank
* Postcode data: IDACI Decile
* Postcode data: IDACI Score
* Postcode data: Income Deprivation Affecting Older People Index (IDAOPI) Rank
* Postcode data: IDAOPI Decile
* Postcode data: IDAOPI Score

*Co-morbidities*

Validated CogStack concepts were used for whole co-morbidity dataset population, with output as 0 (not present) or 1 (present) for following co-morbidities:

* Hypertension
* Heart disease
* Chronic respiratory disease
* Chronic liver disease
* Chronic kidney disease
* Diabetes mellitus
* Atrial fibrillation

Further data was recorded for patients suffering other non-HNCs. Non-HNCs labelled in structured format were taken when extracting diagnosis data (see below). Any missing data was added for patients identified to have had RT for non-HN cancers based on the structured RT data pull from the EDW Mosaiq data tables. This data was recorded in a consistent terminology. Some patients had suffered from 2 separate non-HNC malignancies and data was therefore recorded as:

* Non-HN Cancer Diagnosis 1
* Non-HN Cancer Diagnosis 2

A structured data extract from the EPR data table on the EDW denoting all coded medical issues for all MRNs was subsequently extracted from SSMS. In the future this data can be cleaned and the accuracy of this data could be assessed against our current co-morbidity data from the above methods as a subproject. If robust, this data would not be limited to the existing co-morbidity data categories from CogStack above.

*Disease*

A combination of methods was used to draw together all the relevant diagnosis data for patients. A large amount of structured data was coded on Mosaiq EDW data tables, which was used wherever possible. Other sources of data such as the HNC Masterlist, CogStack output for validated diagnosis concepts, and manually curated data (oral cavity, larynx and oropharynx datasets from previous retrospective research projects) were linked where required. For all relevant data categories, “0” represents a data category to be negative/absent, “1” represents positive/present.

Diagnosis Date

* This data was obtained from the structured diagnosis data extract from the EDW’s Mosaiq diagnosis data table.
* As per the NHS data dictionary (24), this was taken to be the date of the pathology report confirming the primary cancer. If this data was not available, this date was taken as the date of the Multidisciplinary Team Meeting (MDT).
* There were some missing diagnosis dates, and therefore taken as decision to treat date (493 patients).
* If no decision to treat date present, diagnosis dates were manually curated (mainly early 2010 patients – if no histopathology date was available (as histopathology data not commonly recorded on EPR for pre-2014 patients due this being stored in a different system), this date was taken as the first MDT date ratifying diagnosis. If no data was available for either of these dates, diagnosis was then taken as “oncology first visit” date.
* All diagnosis dates were checked with comparison to oncology first visit dates by diagnosis date from oncology first visit date:
  + If negative (oncology first visit occurring before recorded diagnosis date) these were checked manually and amended accordingly (223 dates). Again, these erroneous data points tended to be more historic patient’s ones where diagnosis date had not been recorded appropriately at the time of diagnosis but instead recorded on the initiation of treatment.
  + If the difference was more than 300 days, these dates were also checked: The majority of dates were found to be correct and the difference was due to a disease recurrence (for example post-surgery having not previously seen oncology) now requiring oncology input. Any wrong dates encountered were again corrected.

Decision to treat date

* These dates were also from the structured diagnosis data extract from the EDW’s Mosaiq diagnosis data table. Missing values (137 patients) were not manually curated.

Oncology first visit date

* This date was taken from the HNC Masterlist for all patients, with 45 patients’ dates corrected manually during data curation/verification as above. Differences recorded were mostly no more than 7 days different from actual oncology first visit dates.

Disease major site

* This data was extracted from the HNC Masterlist, and corroborated with the structured diagnosis data extract from the EDW’s Mosaiq diagnosis data table.
* Recorded as one of: Oropharynx, Oral Cavity, Nasopharynx, Hypopharynx, Larynx, Salivary Gland, Paranasal Sinus/Nasal Cavity, CUP, or Multiple.
* To improve accuracy, potentially problematic areas were manually checked:
  + Parotid cancers (to ensure not skin cancer secondary cancers).
  + Maxilla/maxillary sinus cancers (to differentiate oral cavity from paranasal sinus/nasal cavity cancers).
  + Any unclear primary sites due to limited clinical information from the HNC Masterlist.

Multiple HNC details

* For patients with multiple HN cancers identified within the above process, data was manually curated to include disease major site, cTNM7 staging and diagnosis date for each HN cancer.
* All HN cancers were recorded even if one was earlier than 2010 (as this would potentially have affected subsequent treatment of later HN cancers, such as not being able to have further radiotherapy).

Disease subsite

* Data from HNC Masterlist clinical information was amalgamated with data from the pre-existing GT Oropharynx, Oral Cavity and Larynx datasets, as well as when this was clearly coded in the Mosaiq diagnosis data table alongside disease major site.

Disease laterality

* Some data was available from the structured diagnosis data extract from the EDW’s Mosaiq diagnosis data table. This data was combined with laterality data from pre-existing GT Oropharynx and Oral Cavity datasets.
* Laterality was standardised to: Right, Left, Midline, Bilateral, Unknown.

cT stage TNM7, cN stage TNM7, cM stage TNM7, cDisease Stage TNM7

* This data was obtained from the structured diagnosis data extract from the EDW’s Mosaiq diagnosis data table.
* cTNM7 data was not coded on Mosaiq and therefore missing for 458 patients. 61 of these patients’ cTNM7 staging was available and therefore taken from pre-existing GT datasets. The remaining missing data was manually curated given the importance of this data point for completion.
* Full cTNM7 staging was not recorded within any EHR system for 66 patients (labelled as ‘unknown’).

pT stage TNM7, pN stage TNM7, pDisease Stage TNM7

* Data taken from the pre-existing oral cavity GT dataset.
* A limited amount of data points were also manually curated data simultaneously during curation of missing cTNM7 values.

cT stage TNM8, cN stage TNM8, cM stage TNM8, cDisease Stage TNM8

* Data taken from the pre-existing oropharynx and oral cavity GT datasets from previous retrospective project.

pT stage TNM8, pN stage TNM8, pDisease Stage TNM8

* Data taken from the pre-existing oral cavity GT dataset.

Histopathology

* Data extracted from the HNC Masterlist, and corroborated with the structured diagnosis data extract from the EDW’s Mosaiq diagnosis data table.

Histology Code

* Data was obtained from the structured diagnosis data extract from the EDW’s Mosaiq diagnosis data table.

Tumour Grade

* Some data was available from the structured diagnosis data extract from the EDW’s Mosaiq diagnosis data table. This data was combined with laterality data from pre-existing GT oral cavity datasets.

HPV Status

* The pre-existing oropharynx GT datasets contained 524 manually curated values.
* For remaining patients, CogStack output was used.
  + If zero recorded data within a patient’s output for both HPV negative SCC and HPV positive SCC concepts, this was taken as ‘NOT TESTED’.
  + If most output for HPV negative SCC disease taken as HPV negative, if more output for HPV positive SCC disease taken as HPV positive
  + If equal output, disease taken as HPV negative (based on validation project thresholding analysis findings).

HPV p16

* No data currently curated as no structured/existing data available. Aim is to have this data in the future, but will likely require manual curation.

HPV ISH

* No data currently curated as no structured/existing data available. Aim is to have this data in the future, but will likely require manual curation.

EBV Status

* As this data category was only relevant to a limited number of patients, this was manually curated for all nasopharynx cancer patients (118 patients).

Lymphovascular invasion

* CogStack’s performance was not sufficient for automated population of this category. Data present was therefore data from the pre-existing GT oral cavity retrospective dataset.

Perineural invasion present

* CogStack’s performance was adequate to populate this data category – this was combined with a pre-existing GT oral cavity dataset (where discrepancies the GT data was favoured).

Surgical margin

* Data was taken from the pre-existing oral cavity GT datasets from previous retrospective project.

Depth of invasion

* Data was taken from the pre-existing oral cavity GT datasets from previous retrospective project.

Extracapsular extension

* CogStack’s performance was not sufficient for automated population of this category. Data present was therefore data from the pre-existing GT oral cavity retrospective dataset.

Metastases sites at diagnosis

* For patients marked as ‘M1’ in staging, site of metastases was manually curated for lung, brain, liver and/or bone metastases at diagnosis. This was denoted as “1” if metastases at that site at diagnosis, or “0" if either M0 staging or M1 staging but no metastases at that specific site.

PD-L1 score TPS, PD-L1 TPS category, PD-L1 score CPS, PD-L1 CPS category

* All data used was provided by Dr Lisette Collins (Histopathology Consultant) on 29/1/2024 for our centre’s complete PDL1 testing and result data.

*Treatment*

1. *General*

* Initial treatment received
  + Once all complete different treatment modality data (see below) had been amalgamated, initial treatment received was curated by calculating from treatment combinations received and dates of treatment. This data was corroborated with structured data pulls (for example treatment recorded on the EDW’s Mosaiq HN assessment v5 data tables) as wells as HNC Masterlist data.
  + For 400 patients there was still a possibility of inaccuracy from the above approach (for example initial treatment documented by corroborating structured data disagreed with calculated initial treatment from treatment data) and these cases were therefore manually checked and curated to maximise accuracy of this key data category.
* HN list clinical info
* Taken directly from HNC Masterlist without data cleaning.
* Multidisciplinary meeting (MDM) data – all data was from a structured data extraction from the EDW’s Mosaiq HN assessment form data table, providing all EHR documented HN MDM data (May 2012 to October 2023) in chronological order of date of MDT discussion. This data may provide a chronological narrative of a patient’s cancer journey. Each MDM contained the following data:
* MDM date
* MDM Discuss/Outcome 1
* MDM Discuss/Outcome 2
* Treatment Intent
* Treatment Plan
* Other Treatment Plan
* Eligible for Clinical Study
* Clinical Study Name

1. *Surgery*

Surgical data was obtained from a structured data extract from the EDW’s EPR procedure data table for all cohort MRNs. From this extract relevant HN surgical procedures were selected and data was cleaned to remove exact duplicates. If duplicates of the same procedure recorded on different dates were noted, the first record of each procedure was retained. All non-dental procedures for patients where each specific procedure was not recorded was marked as “0” to indicate they had not undergone this procedure. The exception to this was for gastrostomy removal, which likely represented incomplete data as from experience we know that gastrostomy tubes are often removed in the community setting and therefore not coded our centre’s EHR system. Data curated in this way was:

* Laryngectomy
* Laryngectomy Procedure Description
* Laryngectomy Procedure Date
* Neck dissection
* Neck dissection Procedure Description
* Neck dissection Procedure Date
* Tracheostomy
* Tracheostomy Procedure description
* Tracheostomy First Procedure Date
* Mandible excision
* Mandible excision Procedure Description
* Mandible excision Procedure Date
* Maxillectomy
* Maxillectomy Procedure Description
* Maxillectomy Procedure Date
* Total Glossectomy
* Total Glossectomy Procedure Description
* Total Glossectomy Procedure Date
* Partial Glossectomy
* Partial Glossectomy Procedure Description
* Partial Glossectomy Procedure Date
* Oral cavity surgery
* Oral cavity surgery Procedure Description
* Oral cavity surgery Procedure Date
* Tonsillectomy
* Tonsillectomy Procedure Description
* Tonsillectomy Procedure Date
* Pharyngectomy
* Pharyngectomy Procedure Description
* Pharyngectomy Procedure Date
* Salivary gland surgery
* Salivary gland surgery procedure Description
* Salivary gland surgery First procedure Date
* Nasal cavity sinus surgery
* Nasal cavity sinus surgery Procedure Description
* Nasal cavity sinus surgery Procedure Date
* Orbital Exenteration
* Orbital Exenteration Procedure Description
* Orbital Exenteration Procedure Date
* Salvage procedure
* Electrochemotherapy
* Electrochemotherapy procedure Description
* Electrochemotherapy first procedure date
* Gastrostomy
* Gastrostomy procedure Description
* Gastrostomy Date
* Gastrostomy removal
* Gastrostomy removal Procedure Description
* Gastrostomy removal First Procedure Date

CogStack output data was available for a small number of procedures (e.g. laryngectomy) and was included so that discrepancies between CogStack’s output and structured extract surgical data could be checked where required. However based previous CogStack evaluation work and evaluation below, use of the above data categories would be recommended over the CogStack surgical data categories.

1. *Chemotherapy*

All HNC chemotherapy data was obtained from a structured data extract from Mosaiq provided by the Mosaiq team. Data was cleaned by undertaking the following steps:

* Deletion of all less relevant data:
  + Non-systemic anti-cancer treatment (SACT) medication data such as supportive medications.
  + Non-HN SACT data inadvertently included within the data extract.
  + Treatment recorded as prescribed but not given
* Division of data into treatment settings of induction chemotherapy, concomitant chemotherapy and palliative SACT data
* Pivoting of data so each setting contained one row per patient with all data retained

Following this data cleaning process, data was corroborated with CogStack output for various concepts (‘induction chemotherapy’, ‘concurrent chemoradiotherapy’, ‘palliative chemotherapy’) as well as data from HNC Masterlist and treatment plan data from HN assessments v5 forms denoting patients to have received chemotherapy. Any possible missing patients were manually checked (and curated if applicable) to ensure completion of data. Missing patients from initial data pull were: 125 (out of a total of 1056) concomitant chemoradiotherapy patients, 45 (out of 426) induction chemotherapy patients, and 43 (out of 363) palliative SACT patients. Most missing patients were due to:

* Treatment on an EHR (Mosaiq) protocol not pulled into the original data pull e.g. 'Cisplatin +RT (Short hydration protocol)', ‘Cisplatin/Etoposide’ that was linked to a different cancer type (e.g. small cell lung cancer) and therefore not pulled automatically in the extraction provided.
* Treatment prescribed on paper rather on the Mosaiq EHR because of:
  + Treatment during a prolonged IT cut out period (summer 2022)
  + Treatment pre-2012 when chemotherapy Mosaiq prescriptions were introduced
  + Treated at another centre
  + Newer patients after the provided data extraction (treated with SACT between 1/8/23 and 4/10/23)

Final chemotherapy datasets were uploaded to the EDW via the SAM designer tool for linkage with other data categories. The above method was used to curate the following data points:

* Neoadjuvant chemo
* NACTregimen1
* NACT Cycle1 date
* NACT cycles
* Cisplatin NACT cycles
* Cisplatin NACT total dose
* Carboplatin NACT cycles
* Carboplatin NACT total dose
* 5FU NACT doses
* 5FU NACT total dose
* Docetaxel NACT cycles
* Docetaxel NACT total dose
* Gemcitabine NACT cycles
* Gemcitabine NACT total dose
* NACT regimen 2
* Concomitant chemo
* Concomitant drug1
* Concomitant cycle1 date
* Concomitant drug1 total dose
* Concomitant drug1 cycles
* Concomitant drug 1 dosing regimen
* Concomitant drug2
* Concomitant drug2 total dose
* Concomitant drug2 cycles
* Concomitant drug 2 dosing regimen
* HN Palliative SACT
* HN 1st line pall SACT regime
* HN 1st line pall SACT start date
* HN 1st line pall SACT number cycles
* HN 2nd line pall SACT regime
* HN 2nd line pall SACT start date
* HN 2nd line pall SACT number cycles
* HN 3rd line pall SACT regime
* HN 3rd line pall SACT start date
* HN 3rd line pall SACT number cycles
* HN 4th line pall SACT regime
* HN 4th line pall SACT start date
* HN 4th line pall SACT number cycles
* HN 5th line pall SACT regime
* HN 5th line pall SACT start date
* HN 5th line pall SACT number cycles

1. *Radiotherapy*

A structured data extraction of all radiotherapy treatment data on Mosaiq was obtained from the EDW’s Mosaiq EHR treatment data table. Our centre’s RT data recording system changed from Mosaiq to Aria (Eclipse) on 29/9/2021 meaning that patients treated on the latter system were still recorded as having received RT on this data extraction but with other data missing such as RT dose and course date information. Using the SSMS on EDW this data was simultaneously amalgamated with a structured data extraction of Mosaiq RT request form data (provided by the Mosaiq team and uploaded to the EDW using the SAM Designer tool) for HN patients from 2016 (when this request form was first used) until mid-2023 to give further details on initial intention of treatment and patient status (such as whether they are a clinical trial patient or had special RT planning considerations such as the need for an MRI-in-shell). Following amalgamation, data was cleaned to pivot all data to have one row per patient as per our requirements for this dataset. Data required cleaning if patients had multiple rows due to having received multiple different courses of RT or re-plans of one course. Further adjustments were required such as converting units (cGy to Gy). Non-HNC RT courses were included within the data extraction and therefore cleaned and retained in the same way.

Missing data post-29/9/2021 (treatment technique, dose and RT course date data) required manual curation (201 patients). Data regarding early termination of treatment was curated based on identification of these patients by the fractions received differing from the intended fractionation. Other data including pacemaker presence and MRI-in-shell-planning was taken where available from RT request forms. Where missing from RT request form data, treatment intent was reviewed for unclear regimens (such as 55Gy in 20Fr, which could be either palliative or radical intent depending on the context) and manually curated to give complete data for this key data category. If patients did not receive RT, all relevant data points were marked as 'Not applicable'. Incomplete data points were left blank (such as MRI in shell data) that we hope to be completed in future.

Following this data cleaning process, data was corroborated with the existing RT data from imaging within XNAT. Discrepancies were checked with 5 out of 2895 patients requiring amendment (3 had treatment elsewhere therefore hadn’t flagged on data pull, 1 HN course (palliative completing on 4/10/23), 1 rectal cancer RT). The finalised cleaned dataset was uploaded to the EDW using the SAM Designer tool.

A few remaining tasks remain to ensure 100% completion:

* 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.

The following data categories were curated in this way:

* HNCRT
* HNCRT course 1 site
* HNC RT course 1 site details
* HNC RT course 1 intent
* HNC RT course 1 category
* HNC RT course 1 technique
* HNC RT course 1 energy
* HNC RT course 1 prescription point
* HNC RT course 1 dose per fraction
* HNC RT course 1 recorded total dose
* HNC RT course 1 recorded fractions
* HNC RT course 1 initially intended total dose
* HNC RT course 1 initially intended fractions
* HNC RT course 1 start date
* HNC RT course 1 end date
* HNC RT course 1 overall treatment time
* HNC RT course 1 replans
* HNC RT course 1 planning MRI
* HNC RT course 1 interruption
* HNC RT course 1 interruption reason
* HNC RT course 1 early termination
* HNC RT course 1 early termination reason
* HNC RT course 1 High dose volume dose
* HNC RT course 1 Prophylactic dose volume dose
* HNC RT course 1 RT clinical trial
* HNC RT course 1 RT clinical trial details
* HNC RT course 1 Previous relevant RT
* HNC RT course 1 Previous relevant RT details
* HNC RT course 1 pacemaker/implantable cardioverter-defibrillator (ICD)
* HNCRT course 2 site
* HNC RT course 2 intent
* HNC RT course 2 category
* HNC RT course 2 technique
* HNC RT course 2 energy
* HNC RT course 2 prescription point
* HNC RT course 2 dose per fraction
* HNC RT course 2 total dose
* HNC RT course 2 fractions
* HNC RT course 2 start date
* HNC RT course 2 end date
* HNC RT course 2 overall treatment time
* HNC RT course 2 replans
* HNCRT course 3 site
* HNC RT course 3 intent
* HNC RT course 3 category
* HNC RT course 3 technique
* HNC RT course 3 energy
* HNC RT course 3 prescription point
* HNC RT course 3 dose per fraction
* HNC RT course 3 total dose
* HNC RT course 3 fractions
* HNC RT course 3 start date
* HNC RT course 3 end date
* HNC RT course 3 overall treatment time
* HNC RT course 3 replans
* HNCRT course 4 site
* HNCRT course 4 intent
* HNC RT course 4 category
* HNC RT course 4 technique
* HNC RT course 4 energy
* HNC RT course 4 prescription point
* HNC RT course 4 dose per fraction
* HNC RT course 4 total dose
* HNC RT course 4 fractions
* HNC RT course 4 start date
* HNC RT course 4 end date
* HNC RT course 4 overall treatment time
* HNC RT course 4 replans
* HNCRT course 5 site
* HNC RT course 5 intent
* HNC RT course 5 category
* HNC RT course 5 technique
* HNC RT course 5 energy
* HNC RT course 5 prescription point
* HNC RT course 5 dose per fraction
* HNC RT course 5 total dose
* HNC RT course 5 fractions
* HNC RT course 5 start date
* HNC RT course 5 end date
* HNC RT course 5 elapsed time
* HNC RT course 5 replans
* HNCRT course 6 site
* HNC RT course 6 intent
* HNC RT course 6 category
* HNC RT course 6 technique
* HNC RT course 6 energy
* HNC RT course 6 prescription point
* HNC RT course 6 dose per fraction
* HNC RT course 6 total dose
* HNC RT course 6 fractions
* HNC RT course 6 start date
* HNC RT course 6 end date
* HNC RT course 6 overall treatment time
* HNC RT course 6 replans
* HNCRT course 7 site
* HNC RT course 7 intent
* HNC RT course 7 category
* HNC RT course 7 technique
* HNC RT course 7 energy
* HNC RT course 7 prescription point
* HNC RT course 7 dose per fraction
* HNC RT course 7 total dose
* HNC RT course 7 fractions
* HNC RT course 7 start date
* HNC RT course 7 end date
* HNC RT course 7 elapsed time
* HNC RT course 7 replans
* HNCRT course 8 site
* HNC RT course 8 intent
* HNC RT course 8 category
* HNC RT course 8 technique
* HNC RT course 8 energy
* HNC RT course 8 prescription point
* HNC RT course 8 dose per fraction
* HNC RT course 8 total dose
* HNC RT course 8 fractions
* HNC RT course 8 start date
* HNC RT course 8 end date
* HNC RT course 8 overall treatment time
* HNC RT course 8 replans
* NonHNC RT
* NonHN RT course 1 site
* NonHNC RT course 1 intent
* NonHNC RT course 1 category
* NonHNC RT course 1 technique
* NonHNC RT course 1 energy
* NonHNC RT course 1 prescription point
* NonHNC RT course 1 dose per fraction
* NonHNC RT course 1 total dose
* NonHNC RT course 1 fractions
* NonHNC RT course 1 start date
* NonHNC RT course 1 end date
* NonHNC RT course 1 elapsed time
* NonHNC RT course 2 site
* NonHNC RT course 2 intent
* NonHNC RT course 2 category
* NonHNC RT course 2 technique
* NonHNC RT course 2 energy
* NonHNC RT course 2 prescription point
* NonHNC RT course 2 dose per fraction
* NonHNC RT course 2 total dose
* NonHNC RT course 2 fractions
* NonHNC RT course 2 start date
* NonHNC RT course 2 end date
* NonHNC RT course 2 elapsed time

RT treatment planning dose data was also curated. This data was extracted in structured format from ARIA using a custom-made script (in C# coding language), with output exported as a CSV file. Using this script, relevant RT treatment planning dose parameters (structure volume, dmax, dmean, 0.1cc dose, D2, D5, D50, D95, D98) were extracted for OARs and target volumes for all patients with RT treatment data on ARIA (2179 patients). The script encountered errors for a 93 patients and was therefore unable to extract all dose data for certain RT treatments.

Inconsistent labelling of both structure names (e.g. Parotid\_R, RtParotid) and RT treatment plan used (e.g. FINAL, 1FINAL) limited the completeness of data that could be linked to our dataset with assurance of accuracy. Therefore, only data for a patient with a single RT treatment plan name for each structure was retained to avoid incorporation of incorrect dose data (for example for patients who had multiple plans due to replanning). In future we can attempt to match this dose data to the RT imaging data to obtain more comprehensive RT dose data for each patient. Retained data was transformed (using SQL code to pivot), loaded to the EDW using SAM Designer and amalgamated with other data tables. The following RT dose data categories for RT patients were collected:

* GTV\_T\_mean\_dose\_gy
* GTV\_T\_dmax\_gy
* GTV\_T\_d\_point1cc\_gy
* GTV\_T\_volume\_cc
* GTV\_T\_D2\_gy
* GTV\_T\_D5\_gy
* GTV\_T\_D50\_gy
* GTV\_T\_D95\_gy
* GTV\_T\_D98\_gy
* GTV\_Tplus5mm\_mean\_dose\_gy
* GTV\_Tplus5mm\_dmax\_gy
* GTV\_Tplus5mm\_d\_point1cc\_gy
* GTV\_Tplus5mm\_volume\_cc
* GTV\_Tplus5mm\_D2\_gy
* GTV\_Tplus5mm\_D5\_gy
* GTV\_Tplus5mm\_D50\_gy
* GTV\_Tplus5mm\_D95\_gy
* GTV\_Tplus5mm\_D98\_gy
* GTV\_Tplus1cm\_mean\_dose\_gy
* GTV\_Tplus1cm\_dmax\_gy
* GTV\_Tplus1cm\_d\_point1cc\_gy
* GTV\_Tplus1cm\_volume\_cc
* GTV\_Tplus1cm\_D2\_gy
* GTV\_Tplus1cm\_D5\_gy
* GTV\_Tplus1cm\_D50\_gy
* GTV\_Tplus1cm\_D95\_gy
* GTV\_Tplus1cm\_D98\_gy
* GTV\_N\_mean\_dose\_gy
* GTV\_N\_dmax\_gy
* GTV\_N\_d\_point1cc\_gy
* GTV\_N\_volume\_cc
* GTV\_N\_D2\_gy
* GTV\_N\_D5\_gy
* GTV\_N\_D50\_gy
* GTV\_N\_D95\_gy
* GTV\_N\_D98\_gy
* GTV\_Nplus5mm\_mean\_dose\_gy
* GTV\_Nplus5mm\_dmax\_gy
* GTV\_Nplus5mm\_d\_point1cc\_gy
* GTV\_Nplus5mm\_volume\_cc
* GTV\_Nplus5mm\_D2\_gy
* GTV\_Nplus5mm\_D5\_gy
* GTV\_Nplus5mm\_D50\_gy
* GTV\_Nplus5mm\_D95\_gy
* GTV\_Nplus5mm\_D98\_gy
* CTV\_70\_mean\_dose\_gy
* CTV\_70\_dmax\_gy
* CTV\_70\_d\_point1cc\_gy
* CTV\_70\_volume\_cc
* CTV\_70\_D2\_gy
* CTV\_70\_D5\_gy
* CTV\_70\_D50\_gy
* CTV\_70\_D95\_gy
* CTV\_70\_D98\_gy
* CTV\_66\_mean\_dose\_gy
* CTV\_66\_dmax\_gy
* CTV\_66\_d\_point1cc\_gy
* CTV\_66\_volume\_cc
* CTV\_66\_D2\_gy
* CTV\_66\_D5\_gy
* CTV\_66\_D50\_gy
* CTV\_66\_D95\_gy
* CTV\_66\_D98\_gy
* CTV\_65\_mean\_dose\_gy
* CTV\_65\_dmax\_gy
* CTV\_65\_d\_point1cc\_gy
* CTV\_65\_volume\_cc
* CTV\_65\_D2\_gy
* CTV\_65\_D5\_gy
* CTV\_65\_D50\_gy
* CTV\_65\_D95\_gy
* CTV\_65\_D98\_gy
* CTV\_60\_mean\_dose\_gy
* CTV\_60\_dmax\_gy
* CTV\_60\_d\_point1cc\_gy
* CTV\_60\_volume\_cc
* CTV\_60\_D2\_gy
* CTV\_60\_D5\_gy
* CTV\_60\_D50\_gy
* CTV\_60\_D95\_gy
* CTV\_60\_D98\_gy
* CTV\_55\_mean\_dose\_gy
* CTV\_55\_dmax\_gy
* CTV\_55\_d\_point1cc\_gy
* CTV\_55\_volume\_cc
* CTV\_55\_D2\_gy
* CTV\_55\_D5\_gy
* CTV\_55\_D50\_gy
* CTV\_55\_D95\_gy
* CTV\_55\_D98\_gy
* CTV\_54\_mean\_dose\_gy
* CTV\_54\_dmax\_gy
* CTV\_54\_d\_point1cc\_gy
* CTV\_54\_volume\_cc
* CTV\_54\_D2\_gy
* CTV\_54\_D5\_gy
* CTV\_54\_D50\_gy
* CTV\_54\_D95\_gy
* CTV\_54\_D98\_gy
* CTV\_50\_mean\_dose\_gy
* CTV\_50\_dmax\_gy
* CTV\_50\_d\_point1cc\_gy
* CTV\_50\_volume\_cc
* CTV\_50\_D2\_gy
* CTV\_50\_D5\_gy
* CTV\_50\_D50\_gy
* CTV\_50\_D95\_gy
* CTV\_50\_D98\_gy
* CTV\_45\_mean\_dose\_gy
* CTV\_45\_dmax\_gy
* CTV\_45\_d\_point1cc\_gy
* CTV\_45\_volume\_cc
* CTV\_45\_D2\_gy
* CTV\_45\_D5\_gy
* CTV\_45\_D50\_gy
* CTV\_45\_D95\_gy
* CTV\_45\_D98\_gy
* PTV\_70\_mean\_dose\_gy
* PTV\_70\_dmax\_gy
* PTV\_70\_d\_point1cc\_gy
* PTV\_70\_volume\_cc
* PTV\_70\_D2\_gy
* PTV\_70\_D5\_gy
* PTV\_70\_D50\_gy
* PTV\_70\_D95\_gy
* PTV\_70\_D98\_gy
* PTV\_66\_mean\_dose\_gy
* PTV\_66\_dmax\_gy
* PTV\_66\_d\_point1cc\_gy
* PTV\_66\_volume\_cc
* PTV\_66\_D2\_gy
* PTV\_66\_D5\_gy
* PTV\_66\_D50\_gy
* PTV\_66\_D95\_gy
* PTV\_66\_D98\_gy
* PTV\_65\_mean\_dose\_gy
* PTV\_65\_dmax\_gy
* PTV\_65\_d\_point1cc\_gy
* PTV\_65\_volume\_cc
* PTV\_65\_D2\_gy
* PTV\_65\_D5\_gy
* PTV\_65\_D50\_gy
* PTV\_65\_D95\_gy
* PTV\_65\_D98\_gy
* PTV\_60\_mean\_dose\_gy
* PTV\_60\_dmax\_gy
* PTV\_60\_d\_point1cc\_gy
* PTV\_60\_volume\_cc
* PTV\_60\_D2\_gy
* PTV\_60\_D5\_gy
* PTV\_60\_D50\_gy
* PTV\_60\_D95\_gy
* PTV\_60\_D98\_gy
* PTV\_55\_mean\_dose\_gy
* PTV\_55\_dmax\_gy
* PTV\_55\_d\_point1cc\_gy
* PTV\_55\_volume\_cc
* PTV\_55\_D2\_gy
* PTV\_55\_D5\_gy
* PTV\_55\_D50\_gy
* PTV\_55\_D95\_gy
* PTV\_55\_D98\_gy
* PTV\_54\_mean\_dose\_gy
* PTV\_54\_dmax\_gy
* PTV\_54\_d\_point1cc\_gy
* PTV\_54\_volume\_cc
* PTV\_54\_D2\_gy
* PTV\_54\_D5\_gy
* PTV\_54\_D50\_gy
* PTV\_54\_D95\_gy
* PTV\_54\_D98\_gy
* PTV\_50\_mean\_dose\_gy
* PTV\_50\_dmax\_gy
* PTV\_50\_d\_point1cc\_gy
* PTV\_50\_volume\_cc
* PTV\_50\_D2\_gy
* PTV\_50\_D5\_gy
* PTV\_50\_D50\_gy
* PTV\_50\_D95\_gy
* PTV\_50\_D98\_gy
* PTV\_45\_mean\_dose\_gy
* PTV\_45\_dmax\_gy
* PTV\_45\_d\_point1cc\_gy
* PTV\_45\_volume\_cc
* PTV\_45\_D2\_gy
* PTV\_45\_D5\_gy
* PTV\_45\_D50\_gy
* PTV\_45\_D95\_gy
* PTV\_45\_D98\_gy
* RT\_Course\_Name
* RT\_Plan\_Name
* Parotid\_R\_volume\_cc
* Parotid\_R\_mean\_dose\_gy
* Parotid\_R\_dmax\_gy
* Parotid\_R\_d\_point1cc\_gy
* Parotid\_L\_volume\_cc
* Parotid\_L\_mean\_dose\_gy
* Parotid\_L\_dmax\_gy
* Parotid\_L\_d\_point1cc\_gy
* PharynxConst\_I\_mean\_dose\_gy
* PharynxConst\_I\_dmax\_gy
* PharynxConst\_I\_d\_point1cc\_gy
* PharynxConst\_I\_volume\_cc
* PharynxConst\_M\_mean\_dose\_gy
* PharynxConst\_M\_dmax\_gy
* PharynxConst\_M\_d\_point1cc\_gy
* PharynxConst\_M\_volume\_cc
* PharynxConst\_S\_mean\_dose\_gy
* PharynxConst\_S\_dmax\_gy
* PharynxConst\_S\_d\_point1cc\_gy
* PharynxConst\_S\_volume\_cc
* OralCavity\_mean\_dose\_gy
* OralCavity\_dmax\_gy
* OralCavity\_d\_point1cc\_gy
* OralCavity\_volume\_cc
* Larynx\_mean\_dose\_gy
* Larynx\_dmax\_gy
* Larynx\_d\_point1cc\_gy
* Larynx\_volume\_cc
* SpinalCord\_PRV\_mean\_dose\_gy
* SpinalCord\_PRV\_dmax\_gy
* SpinalCord\_PRV\_d\_point1cc\_gy
* SpinalCord\_PRV\_volume\_cc
* SpinalCord\_mean\_dose\_gy
* SpinalCord\_dmax\_gy
* SpinalCord\_d\_point1cc\_gy
* SpinalCord\_volume\_cc
* Brainstem\_PRV\_mean\_dose\_gy
* Brainstem\_PRV\_dmax\_gy
* Brainstem\_PRV\_d\_point1cc\_gy
* Brainstem\_PRV\_volume\_cc
* Brainstem\_mean\_dose\_gy
* Brainstem\_dmax\_gy
* Brainstem\_d\_point1cc\_gy
* Brainstem\_volume\_cc
* Brain\_mean\_dose\_gy
* Brain\_dmax\_gy
* Brain\_d\_point1cc\_gy
* Brain\_volume\_cc
* Submandibular\_R\_mean\_dose\_gy
* Submandibular\_R\_dmax\_gy
* Submandibular\_R\_d\_point1cc\_gy
* Submandibular\_R\_volume\_cc
* Submandibular\_L\_mean\_dose\_gy
* Submandibular\_L\_dmax\_gy
* Submandibular\_L\_d\_point1cc\_gy
* Submandibular\_L\_volume\_cc
* Lens\_L\_mean\_dose\_gy
* Lens\_L\_dmax\_gy
* Lens\_L\_d\_point1cc\_gy
* Lens\_L\_volume\_cc
* Lens\_R\_mean\_dose\_gy
* Lens\_R\_dmax\_gy
* Lens\_R\_d\_point1cc\_gy
* Lens\_R\_volume\_cc
* OpticNerve\_L\_mean\_dose\_gy
* OpticNerve\_L\_dmax\_gy
* OpticNerve\_L\_d\_point1cc\_gy
* OpticNerve\_L\_volume\_cc
* OpticNerve\_R\_mean\_dose\_gy
* OpticNerve\_R\_dmax\_gy
* OpticNerve\_R\_d\_point1cc\_gy
* OpticNerve\_R\_volume\_cc
* OpticChiasm\_mean\_dose\_gy
* OpticChiasm\_R\_dmax\_gy
* OpticChiasm\_d\_point1cc\_gy
* OpticChiasm\_volume\_cc
* Retina\_L\_mean\_dose\_gy
* Retina\_L\_dmax\_gy
* Retina\_L\_d\_point1cc\_gy
* Retina\_L\_volume\_cc
* Retina\_R\_mean\_dose\_gy
* Retina\_R\_dmax\_gy
* Retina\_R\_d\_point1cc\_gy
* Retina\_R\_volume\_cc

Data was also taken from a structured data extraction of all HN target volume forms provided by the Mosaiq team. This data was from an EHR assessment form originally completed by the treating oncologist for each RT course. The structured data extraction from the Mosaiq team was uploaded to the EDW via the SAM Designer Tool for inclusion in the dataset, giving the following data points:

* HN Site
* HN Planning System
* HN Planning CT Date
* HN Laterality
* PTV1 Dose
* PTV2 Dose
* PTV3 Dose
* Fractionation
* HN GTV
* HN GTV Rt Nodal Levels
* HN GTV Lt Nodal Levels
* HN CTV1 Organ T
* HN CTV1 Rt Nodal Levels
* HN CTV1 Lt Nodal Levels
* HN CTV2 Organ T
* HN CTV2 Rt Nodal Levels
* HN CTV2 Lt Nodal Levels
* HN CTV3 Tumour Organ
* HN CTV3 Rt Nodal Levels
* HN CTV3 Lt Nodal Levels
* CTV1 to PTV1 margin mm
* HN CTV2 to PTV2
* HN CTV3 to PTV3
* HN GTVT Vol
* HN GTVN Vol
* HN CTV1 Radical Vol
* HN CTV2 Elective Vol
* HN CTV3 Other Vol
* HN PTV1 Radical Vol
* HN PTV2 Elective Vol
* HN PTV3 Other Vol
* HN Spinal Cord Vol
* HN Brainstem Vol
* HN Parotid Rt Vol
* HN Parotid Lt Vol
* HN OpticNerve Rt Vol
* HN OpticNerve Lt Vol
* HN Optic Chiasm Vol
* HN Lens Rt Vol
* HN Lens Lt Vol
* HN Retina Rt Vol
* HN Retina Lt Vol
* HN Pituitary Vol
* HN Cochlea Rt Vol
* HN Cochlea Lt Vol
* HN Accepted Deviation
* HN Vols Completed
* Assessment Date

1. *Dental*

Dental data was obtained from a structured data extract from the EDW’s EPR procedure data table for all cohort MRNs in the same way as the surgical procedure data. From this extract relevant dental procedures were selected and data was cleaned to remove exact duplicates, recording multiple dental extractions on different dates as new columns when data was pivoted to one row per patient. Data was curated in this way for the following data categories:

* Surgical Removal of Tooth Not Elsewhere Classified (NEC)
* Surgical Removal of Tooth Not Elsewhere Classified Date
* Surgical Removal of Tooth Not Elsewhere Classified 2
* Surgical Removal of Tooth Not Elsewhere Classified 2 Date
* Surgical Removal of Wisdom Tooth NEC
* Surgical Removal of Wisdom Tooth NEC Date
* Surgical Removal of Wisdom Tooth NEC\_2
* Surgical Removal of Wisdom Tooth NEC Date \_2
* Extraction of Multiple Teeth NEC
* Extraction of Multiple Teeth NEC Date
* Extraction of Multiple Teeth NEC\_2
* Extraction of Multiple Teeth NEC Date \_2
* Extraction of Multiple Teeth NEC\_3
* Extraction of Multiple Teeth NEC Date \_3
* Surgical Removal of Retained Root of Tooth
* Surgical Removal of Retained Root of Tooth Date
* Full Dental Clearance
* Full Dental Clearance Date
* Lower Dental Clearance
* Lower Dental Clearance Date
* Upper Dental Clearance
* Upper Dental Clearance Date
* Surgical Removal of Impacted Wisdom Tooth
* Surgical Removal of Impacted Wisdom Tooth Date
* Surgical Removal of Impacted Wisdom Tooth\_2
* Surgical Removal of Impacted Wisdom Tooth\_2 Date
* Unspecified surgical removal of tooth
* Unspecified surgical removal of tooth Date
* Unspecified simple extraction of tooth
* Unspecified simple extraction of tooth Date
* Other specified simple extraction of tooth
* Other specified simple extraction of tooth Date
* Surgical Removal of Impacted Tooth Not Elsewhere Classified
* Surgical Removal of Impacted Tooth Not Elsewhere Classified Date

*Bloods*

Haemoglobin (HB), white cell (WCC), lymphocyte, neutrophil, platelet counts, blood group, albumin, alkaline phosphatase (ALP), alanine transferase (ALT), creatinine, corrected calcium, bilirubin were obtained from a structured data extract from the EDW’s EPR result table for all cohort MRNs. Data was then cleaned:

* Haemoglobin values were transformed to convert all values to units of mg/dL (changed on EPR in May 2013 from g/dL).
* All NULL values (clotted/failed samples) were deleted.

Blood result data dates were compared to diagnosis date. If missing (as this was done pre-finalisation of diagnosis dates) this was altered to decision to treat date, and if that value was missing this date was set as the oncology first visit date. As this was done prior to subsequent cleaning of diagnosis dates, the blood result at diagnosis date may not be accurate for approximately 10% of cohort’s patients but these results can be easily identified. All results for other time points of interest (oncology first visit, RT start, RT end) accord to the final dataset. The difference between each date point and the date of blood results was calculated and the blood result retained was the value closest to the date of interest in question (for example oncology first visit date) for each patient. If no bloods were within 31 days of the relevant date (1 month), this data point was taken as never recorded.

The following data categories were recorded using the above method:

* BloodGroup
* HB DxResultDate
* HB DxResultValue
* HB FirstVisitResultDate
* HB FirstVisitResultValue
* HB RTStartResultDate
* HB RTStartResultValue
* HB RTEndResultDate
* HB RTEndResultValue
* MCV DxResultDate
* MCV DxResultValue
* MCV FirstVisitResultDate
* MCV FirstVisitResultValue
* MCV RTStartResultDate
* MCV RTStartResultValue
* MCV RTEndResultDate
* MCV RTEndResultValue
* WBC DxResultDate
* WBC DxResultValue
* WBC FirstVisitResultDate
* WBC FirstVisitResultValue
* WBC RTStartResultDate
* WBC RTStartResultValue
* WBC RTEndResultDate
* WBC RTEndResultValue
* Neutrophils DxResultDate
* Neutrophils DxResultValue
* Neutrophils FirstVisitResultDate
* Neutrophils FirstVisitResultValue
* Neutrophils RTStartResultDate
* Neutrophils RTStartResultValue
* Neutrophils RTEndResultDate
* Neutrophils RTEndResultValue
* Lymphocytes DxResultDate
* Lymphocytes DxResultValue
* Lymphocytes FirstVisitResultDate
* Lymphocytes FirstVisitResultValue
* Lymphocytes RTStartResultDate
* Lymphocytes RTStartResultValue
* Lymphocytes RTEndResultDate
* Lymphocytes RTEndResultValue
* Platelets DxResultDate
* Platelets DxResultValue
* Platelets FirstVisitResultDate
* Platelets FirstVisitResultValue
* Platelets RTStartResultDate
* Platelets RTStartResultValue
* Platelets RTEndResultDate
* Platelets RTEndResultValue
* Albumin DxResultDate
* Albumin DxResultValue
* Albumin FirstVisitResultDate
* Albumin FirstVisitResultValue
* Albumin RTStartResultDate
* Albumin RTStartResultValue
* Albumin RTEndResultDate
* Albumin RTEndResultValue
* ALP DxResultDate
* ALP DxResultValue
* ALP FirstVisitResultDate
* ALP FirstVisitResultValue
* ALP RTStartResultDate
* ALP RTStartResultValue
* ALP RTEndResultDate
* ALP RTEndResultValue
* ALT DxResultDate
* ALT DxResultValue
* ALT FirstVisitResultDate
* ALT FirstVisitResultValue
* ALT RTStartResultDate
* ALT RTStartResultValue
* ALT RTEndResultDate
* ALT RTEndResultValue
* Bili DxResultDate
* Bili DxResultValue
* Bili FirstVisitResultDate
* Bili FirstVisitResultValue
* Bili RTStartResultDate
* Bili RTStartResultValue
* Bili RTEndResultDate
* Bili RTEndResultValue
* Corrected Calcium DxResultDate
* Corrected Calcium DxResultValue
* Corrected Calcium FirstVisitResultDate
* Corrected Calcium FirstVisitResultValue
* Corrected Calcium RTStartResultDate
* Corrected Calcium RTStartResultValue
* Corrected Calcium RTEndResultDate
* Corrected Calcium RTEndResultValue
* Creatinine DxResultDate
* Creatinine DxResultValue
* Creatinine FirstVisitResultDate
* Creatinine FirstVisitResultValue
* Creatinine RTStartResultDate
* Creatinine RTStartResultValue
* Creatinine RTEndResultDate
* Creatinine RTEndResultValue

*Outcomes*

A structured data extract obtained from the EDW’s PIMS patient data table for last attended follow date (before death date if had died), location and type, as well as alive/death status and death date. If a patient had died during an inpatient admission at our centre, death date was recorded as the same date as last follow-up. Overall survival time and status at different time points can easily be calculated using this data. Data categories collected in the above way were:

* Last follow-up date
* Alive/dead at last follow-up date
* Date of death

CogStack output was used for:

* Brain metastases at any point
* CogStack\_CR\_and\_radical\_RT (patients who CogStack detected as having a complete response (no temporality given), known to have had radical RT)

Limited structured EHR data was available for remaining outcome data categories. Some existing data from GT datasets was available and used, and some data could be taken from the HN assessment forms ‘disease assessment’ status. In combination with this, further data could be inferred (for example, which data categories weren’t applicable for which patients – for example if they never received treatment (had best supportive care), disease response/failure was not applicable), or from neck dissection dates following radical RT for salvage neck dissection data. Using the above combined approach, a large amount of data was curated, however there remains a significant amount of data that will likely require manual curation in the future as no structured data extract was available. The following data categories were recorded using the above method:

* 3 month RT response
* 3 month imaging modality
* 3 month imaging date
* 6 month RT response
* 6 month imaging modality
* 6 month imaging date
* Biopsy post-RT
* Biopsy date
* Salvage neck dissection
* Salvage neck dissection date
* Cause of death
* Cause of death - other
* Failure
* Failure date
* Failure site summary
* Primary failure site details
* Primary recurrence date
* Primary recurrence Rx
* Primary recurrence intent
* Nodal or locoregional failure site
* Nodal or locoregional non-primary recurrence date
* Nodal recurrence Rx
* Nodal recurrence intent
* Metastatic site
* Met recurrence date
* Met recurrence Rx
* Met recurrence intent
* Recurrence after RT
* Time to failure after RT
* In field recurrence high dose
* In field recurrence ppx dose

Discussions were held with NHS England about obtaining death data (cause of death and death date) for our cohort from NCRAS, however this would have to be subject to obtaining several governance permissions and would have cost in excess of £20000 for 3 years of data access, and was therefore not pursued further.

*CTCAE toxicity*

A structured data pull from the EDW’s Mosaiq HN assessment v5 data table was extracted for CTCAE toxicity data. This data comprised of structured data forms completed on each patient attendance to the HNC outpatient clinic. This form was first put into place in November 2011 therefore data was not present for NCI CTCAE toxicity data points before this point. Data recorded on this form included NCI CTCAE (v4 data up until 20 October 2020, v5 data after 20 October 2020) and RTOG data. Whilst data is recorded with an assessment point (for example ‘1 year post-completion of radiotherapy’), this completed assessment point was known from experience to not always be accurate (due to incorrect recording often related to changes to patients out patients appointments) therefore RT start and end dates were aggregated with this data to allow accuracy of assessment completion. By calculating the time difference (between RT start date and assessment form for on treatment assessment forms, and RT end date and assessment for post-treatment assessment forms), accuracy of timepoint of toxicity documentation was assured. Data was cleaned by removing blank and duplicate forms. If there were multiple completed forms for a single assessment time point, the most complete data was selected. Date ranges were used to ensure capture of as much relevant CTCAE data as possible. These date ranges were:

* 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

Data was subsequently pivoted to one row per patient. Post-treatment date ranges could be adjusted in the future to include any assessment forms not included as between chosen date ranges. The following data categories were present in the final dataset for each of the above assessment points (RTOG grading data categories incorporated in baseline and then 3 month post-RT assessments onwards, as these solely assessed late effects of RT):

* Toxicity assessment date (retained allowing ascertainment of whether CTCAE data present relating to CTCAE v4 or v5)
* Aspiration CTCAE grading
* Dysphagia CTCAE grading
* Facial oedema CTCAE grading
* Mucositis CTCAE grading
* Nature of diet
* Dysphonia CTCAE grading
* Pharyngolaryngeal pain CTCAE grading
* Radiation dermatitis CTCAE grading
* Tinnitus CTCAE grading
* Trismus CTCAE grading
* Xerostomia CTCAE grading
* Feeding tube dependence
* Osteoradionecrosis RTOG grading
* Salivary gland toxicity RTOG grading
* Skin toxicity RTOG grading
* Speech RTOG grading
* Clinical study
* Disease assessment
* Treatment intent
* Treatment type
* Palliative chemo receiving
* Weight

*Patient Reported Outcome Measures (PROMS, EORTC and QLQ data)*

PROMs data had been recorded for patients attending the HNC oncology clinic since October 2021 for EORTC H&N43 data and January 2023 for QLQ-C30. No PROMS data was recorded before October 2021. This data was originally recorded for the vast majority of patients electronically on the DrDoctor online application. Data from this was extracted in CSV file format on 5/10/23, uploaded to the EDW and pivoted using SSMS to ensure one row per patient with all data present in chronological order. It was noted for categories where ‘not applicable’ is an option, scores may be calculated to a negative score (e.g. swallowing, speech, social eating, and sexuality) on DrDoctor, which is not possible (scoring system is 0 to 100). These negative scores were therefore converted to ‘Not applicable’ to account for this.

The following data categories were included within the dataset:

* *EORTC* *H&N43 data*
* Form Name
* Form Completed Date
* Calculated Pain in mouth
* Calculated Swallowing
* Calculated Teeth
* Calculated Dry Mouth
* Calculated Problems with senses
* Calculated Speech
* Calculated Body Image
* Calculated Social Eating
* Sexuality
* Calculated Shoulder
* Calculated Skin Problems
* Calculated Fear of progression
* Calculated Mouth Opening
* Calculated Cough
* Calculated Social contact
* Calculated Swelling in the neck
* Calculated Weight loss
* Calculated Problems with wound healing
* Calculated Neurological problems
* *QLQ-C30 data*
  + Form Completed Date
  + Calculated QLQ C30 Score 1 QL2
  + Calculated QLQ C30 Score 2 PF2
  + Calculated QLQ C30 Score 3 RF2
  + Calculated QLQ C30 Score 4 EF
  + Calculated QLQ C30 Score 5 CF
  + Calculated QLQ C30 Score 6 SF
  + Calculated QLQ C30 Score 7 FA
  + Calculated QLQ C30 Score 8 NV
  + Calculated QLQ C30 Score 9 PA
  + Calculated QLQ C30 Score 10 DY
  + Calculated QLQ C30 Score 11 SL
  + Calculated QLQ C30 Score 12 AP
  + Calculated QLQ C30 Score 13 CO
  + Calculated QLQ C30 Score 14 DI
  + Calculated QLQ C30 Score 15 FI

*On-treatment support data*

Structured data extracts had not been obtained at the time of writing for most treatment support data categories at the time of writing, but we hope to curate this data in the future. These will require a combination of structured data pulls (where available) from legacy system tables on EDW and manual curation. Data categories of interest are:

* Feeding tube recommendation
* Prophylactic/reactive
* Tube type (nasogastric tube (NGT)/percutaneous endoscopic gastrostomy (PEG)/radiologically-inserted gastrostomy (RIG))
* Feeding tube inserted date
* Feeding tube use at the start of radiotherapy
* Feeding tube use at the end of radiotherapy
* Feeding tube removal date
* Duration of feeding tube in situ
* Hospital admission
* Reason for admission

*Imaging study data*

Full documentation of all imaging studies (including both diagnostic and radiotherapy imaging data such as planning CT scans, structure contours and dose map data) available on XNAT for each patient in this cohort can be found in the relevant imaging SOP documentation (Document 011 Transfer of Imaging Data SOP and Document 013 Transfer of Radiotherapy Data SOP).

**Amalgamation of data**

All finalised data tables were uploaded to the EDW using SAM Designer. SQL code was then written to join all relevant data categories from different tables on SSMS, to form the complete retrospective dataset which could be copied to a CSV file (in the format of one row per patient containing all data categories as above) for downstream use, alongside relevant imaging data.

The final code combined 100 different data tables, primarily joined on patient MRN. This code length exceeded the memory capacity of our centre’s version of SSMS for a single query and therefore was split into 3 separate codes (one code of 655 lines, a second code of 1238 lines, and a third code of 286 lines – finalised code can be seen in Appendix 2), the output of which could then be copied and amalgamated on a CSV file. For completeness, this code contained all data categories available.

**Data validation**

To ensure reliability and relevance of our clinical dataset, dataset quality was assessed across the data quality dimensions of accuracy, completeness, timeliness and consistency. As per published guidance (3), this process was performed both during the dataset build but also a final validation assessment was completed for the complete finalised dataset amalgamated using the above methodology against the relevant data quality dimensions. Following data validation, any incorrect data found were corrected within the dataset. Data uniqueness was not formally re-assessed on completion as during the clinical dataset build the dataset was designed in such a way as to ensure no duplication of data categories within the dataset.

*Data accuracy and timeliness assessment*

Given the size of the dataset, it was not feasible to assess large amounts of data against manually checked data, and therefore a subsample was checked, with prioritisation to key data categories as recommended within the literature (3). Microsoft Excel’s randomiser function was used to select a random sample of 60 patients for each data category for assessment. To ensure as comprehensive assessment of the retrospective dataset was performed as practically possible, a different random sample was selected for each of the key data subdivisions (demographics, comorbidities, disease factors, overall treatment, radiotherapy, surgery, chemotherapy, outcomes). For each patient randomly selected, the data category was manually assessed by reviewing the EHR and the recorded data point compared to the data point recorded within the final clinical dataset. A recording of “1” (correct) or “0” (incorrect) was made. From this point data accuracy could be estimated for each data category. For binary categories positive predictive value (PPV, True Positives/(True Positives + False Positives)) and negative predictive value (NPV, True Negatives/(True Negatives + False Negatives)) were also estimated. For timeliness, the timepoint of relevance taken for assessment against was 4/10/2023 for the purposes of this project.

For data categories that were direct structured EHR data pulls that did not undergo any data cleaning or transformation, a sample of 20 for each (MDM, bloods, CTCAE, PROMS, RT dose data) was assessed to ensure this data was entirely accurate as reflected in the original recorded structured data source.

*Data completeness assessment*

Data completeness was calculated for each data category on Microsoft Excel by using the COUNTIF function to calculate complete and incomplete data points (‘NULL’ (the value all blank data points appear on the EDW), ‘Not stated’, ‘Unknown’). From this the percentage of data completion for each data category could then be calculated. Consideration for each data category was given to the fact that incomplete data did not necessarily mean missing data depending on the data category. For example, a missing TNM staging would likely represent genuine missing data as all cancer patients should undergo TNM staging, whereas a missing 1 year post-RT assessment form may have meant that this assessment could never have happened as a patient never received RT or died before 1 year post-radiotherapy completion. For some RT data such RT target volumes and RT dose data, completeness was calculated against the number of patients who received RT at GSTT (2553) rather than the entire population.

*Data consistency assessment*

Given the large amount of data categories, it was not feasible to go through every single category and record consistency quantitatively. This was therefore assessed for each data category as above during the dataset build and in the final validation process to ensure data was recorded in a standardised and consistent manner.

**Permissions**

Permission to undertake this work was approved by our centre’s Information Governance team following completion of a Data Protection Information Assessment covering all aspects of this work both imaging (using XNAT as the research database platform of choice) and clinical data curation. Patient data was used as per the opt-out consent process underpinned by our centre’s Research Ethics Committee Guy’s Cancer Cohort ethics approval process (REC Reference:18/NW/0297, IRAS Project ID: 231443) (26). Appendix 3 shows all those involved in the development of this retrospective clinical dataset that may require citation depending on what data is used in downstream projects.

**Results**

**Final clinical dataset**

The final retrospective cohort consisted of 2895 patients who met the inclusion criteria for the clinical dataset. In total, over 1.7 million data points were captured within the dataset (Table 1). A full breakdown of data is beyond the scope of this SOP, but examples of the population breakdown can be seen in Appendix 4.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Patients** | **Data categories** | **Complete data points** | **Unknown/NULL/Not stated data points** |
| **Total** | 2895 | 2077 | 1803670 | 4212494 |

Table 1. Overview of clinical dataset characteristics (correct as of 11/11/24). Unknown/NULL/Not stated data points includes data points where data may not be expected (for example, 2 year toxicity scoring for a patient that died 1 year after radiotherapy), rather than a true value on ‘missing’ data not yet curated.

**Data validation**

Data quality was assessed across different data quality dimensions to assess the relevance and reliability of the clinical dataset built using the methodology set out above.

*Data accuracy and timeliness assessment*

By assessing a sample of data present within the final clinical dataset from 60 randomly selected patients against manually curated data for each subdivision of demographics, comorbidities, disease factors, overall treatment, radiotherapy, surgery, chemotherapy, and outcome data, and estimation of dataset accuracy was calculated. For data categories, data was assessed for all patients where data was present (for example, all patients had recorded postcode, therefore this was recorded as correct or incorrect) but if data was present for only a proportion of the 60 patients (for example alcohol intake amount), only the data present was assessed for accuracy.

For demographics, all demographic data within the sample was found to be 100% accurate apart from post code data, where 1 patient had changed address but post code on all EHRs had not been changed (Table 2).

|  |  |  |  |
| --- | --- | --- | --- |
| **Demographic data category** | **Number assessed** | **Number correct** | **% correct** |
| DOB | 60 | 60 | 100.0% |
| Gender | 60 | 60 | 100.0% |
| Ethnicity | 60 | 60 | 100.0% |
| Postcode | 60 | 59 | 98.3% |
| Address | 60 | 59 | 98.3% |
| Height\_cm | 9 | 9 | 100.0% |
| First\_recorded\_weight\_kg | 10 | 10 | 100.0% |
| Date\_weight\_recorded | 10 | 10 | 100.0% |
| ECOG\_Performance\_Status\_ObservationDTS | 10 | 10 | 100.0% |
| First\_recorded\_ECOG\_Performance\_Status | 10 | 10 | 100.0% |
| Alcohol\_Observation\_DTS | 10 | 10 | 100.0% |
| Alcohol\_Observation\_Choice\_ShortNM | 10 | 10 | 100.0% |
| Alcohol\_Amount\_ObservationDTS | 3 | 3 | 100.0% |
| Alcohol\_Amount\_ObservationChoiceShortNM | 3 | 3 | 100.0% |
| Smoking\_observation\_date | 10 | 10 | 100.0% |
| First\_recorded\_Smoking\_status | 10 | 10 | 100.0% |
| Smoking\_Amount\_ObservationDTS | 2 | 2 | 100.0% |
| Smoking\_Amount\_ObservationChoiceShortNM | 2 | 2 | 100.0% |

Table 2. Accuracy estimation results for demographic data categories.

Co-morbidity data was in excess of 95% for all categories, with the joint lowest performing categories being hypertension, heart disease, liver disease diabetes mellitus, non-HNC cancer diagnosis (all 96.7% correct), with 100% accuracy seen for atrial fibrillation and chronic respiratory disease (Table 3).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **PMH data category** | **Number assessed** | **Number correct** | **% correct** | **PPV** | **NPV** |
| Hypertension | 60 | 58 | 96.7% | 0.952 | 0.974 |
| Heart\_disease | 60 | 58 | 96.7% | Incalculable | 0.967 |
| Chronic\_resp\_disease | 60 | 60 | 100.0% | 1.000 | 1.000 |
| Chronic\_liver\_disease | 60 | 58 | 96.7% | 1.000 | 0.965 |
| Chronic\_kidney\_disease | 60 | 59 | 98.3% | 1.000 | 0.983 |
| Diabetes\_mellitus | 60 | 58 | 96.7% | 1.000 | 0.966 |
| Atrial\_fibrillation | 60 | 60 | 100.0% | 1.000 | 1.000 |
| NonHNCCancerDx1 | 60 | 58 | 96.7% | 1.000 | 0.963 |
| NonHNCCancerDx2 | 60 | 60 | 100.0% | Incalculable | 1.000 |

Table 3. Accuracy, PPV and NPV estimation results for co-morbidity data categories.

Disease characteristic data was estimated to be 95% accurate for all categories except disease laterality and perineural invasion (Table 4). 15 categories were estimated to be 100% correct including diagnosis date, oncology first visit and major disease site. Initial treatment received was estimated to be 96.7% accurate (2 patients received concomitant chemotherapy, not previously picked up as this had been given on a paper prescription (both patients were treated pre-2012)).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Dx data category** | **Number assessed** | **Number correct** | **% correct** | **PPV** | **NPV** |
| Diagnosis\_Date\_Mosaiq | 60 | 60 | 100.0% |  |  |
| Onc\_first\_visit | 60 | 60 | 100.0% |  |  |
| Disease\_major\_site\_HNlist | 60 | 60 | 100.0% |  |  |
| Multiple\_Sites\_Details | 60 | 60 | 100.0% |  |  |
| Disease\_subsite | 25 | 25 | 100.0% |  |  |
| Disease\_laterality\_Mosaiq | 29 | 27 | 93.1% |  |  |
| cT\_stage\_Mosaiq\_TNM7 | 60 | 58 | 96.7% |  |  |
| cN\_stage\_Mosaiq\_TNM7 | 60 | 57 | 95.0% |  |  |
| cM\_stage\_Mosaiq\_TNM7 | 60 | 58 | 96.7% |  |  |
| cDisease\_Stage\_TNM7 | 60 | 57 | 95.0% |  |  |
| Lung\_mets\_at\_Dx\_CogStack | 48 | 48 | 100.0% | 1 | 1 |
| Histopathology | 60 | 60 | 100.0% |  |  |
| Tumour\_Grade\_Mosaiq | 6 | 6 | 100.0% |  |  |
| HPV\_Status\_CogStack | 60 | 60 | 100.0% | 1 | 1 |
| EBV\_Status | 60 | 60 | 100.0% | 1 | 1 |
| Perineural\_invasion\_present\_CogStack | 20 | 17 | 85.0% | 0.5 | 1 |
| Surgical\_margin | 5 | 5 | 100.0% |  |  |
| Extracapsular\_extension\_of nodal\_tumor | 5 | 5 | 100.0% |  |  |
| Lymphovascular invasion | 5 | 5 | 100.0% |  |  |
| PDL1\_score\_TPS | 60 | 60 | 100.0% |  |  |
| PDL1\_TPS\_category | 60 | 60 | 100.0% |  |  |
| PDL1\_score\_CPS | 60 | 60 | 100.0% |  |  |
| PDL1\_CPS\_category | 60 | 60 | 100.0% |  |  |
| Initial\_treatment\_received | 60 | 58 | 96.7% |  |  |

Table 4. Accuracy estimation results for disease characteristic and first treatment data categories.

Chemotherapy data was estimated to be more than 95% accurate for all categories. Of all chemotherapy data, only 1 patient was found to have any incorrect data (Table 5). This pre-2012 patient had been recorded as not receiving neoadjuvant chemotherapy wrongly (as was treatment was prescribed on paper). PPV and NPV was 1 for neoadjuvant, concomitant and palliative SACT binary categories.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Chemo data category** | **Number assessed** | **Number correct** | **% correct** | **PPV** | **NPV** |
| Neoadjuvant\_chemo | 60 | 60 | 100.0% | 1 | 1 |
| NACTregimen1 | 60 | 59 | 98.3% |  |  |
| NACT\_Cycle1\_date | 60 | 59 | 98.3% |  |  |
| NACT\_cycles | 60 | 59 | 98.3% |  |  |
| Cisplatin\_NACT\_cycles | 60 | 59 | 98.3% |  |  |
| Cisplatin\_NACT\_total\_dose | 60 | 59 | 98.3% |  |  |
| Carboplatin\_NACT\_cycles | 60 | 59 | 98.3% |  |  |
| Carboplatin\_NACT\_total\_dose | 60 | 60 | 100.0% |  |  |
| 5FU\_NACT\_doses | 60 | 59 | 98.3% |  |  |
| 5FU\_NACT\_total\_dose | 60 | 59 | 98.3% |  |  |
| Docetaxel\_NACT\_cycles | 60 | 60 | 100.0% |  |  |
| Docetaxel\_NACT\_total\_dose | 60 | 60 | 100.0% |  |  |
| Gemcitabine\_NACT\_cycles | 60 | 60 | 100.0% |  |  |
| Gemcitabine\_NACT\_total\_dose | 60 | 60 | 100.0% |  |  |
| NACT\_regimen\_2 | 60 | 60 | 100.0% |  |  |
| Concomitant\_chemo | 60 | 60 | 100.0% | 1 | 1 |
| Concomitant\_drug1 | 60 | 60 | 100.0% |  |  |
| Concomitant\_cycle1\_date | 60 | 60 | 100.0% |  |  |
| Concomitant\_drug1\_total\_dose | 60 | 60 | 100.0% |  |  |
| Concomitant\_drug1\_cycles | 60 | 60 | 100.0% |  |  |
| Concomitant\_drug\_1\_dosing\_regimen | 60 | 60 | 100.0% |  |  |
| Concomitant\_drug2 | 60 | 60 | 100.0% |  |  |
| Concomitant\_drug2\_total\_dose | 60 | 60 | 100.0% |  |  |
| Concomitant\_drug2\_cycles | 60 | 60 | 100.0% |  |  |
| Concomitant\_drug\_2\_dosing\_regimen | 60 | 60 | 100.0% |  |  |
| HN\_Palliative\_SACT | 60 | 60 | 100.0% | 1 | 1 |
| HN\_1st\_line\_pall\_SACT\_regime | 60 | 60 | 100.0% |  |  |
| HN\_1st\_line\_pall\_SACT\_start\_date | 60 | 60 | 100.0% |  |  |
| HN\_1st\_line\_pall\_SACT\_number\_cycles | 60 | 60 | 100.0% |  |  |
| HN\_2nd\_line\_pall\_SACT\_regime | 60 | 60 | 100.0% |  |  |
| HN\_2nd\_line\_pall\_SACT\_start\_date | 60 | 60 | 100.0% |  |  |
| HN\_2nd\_line\_pall\_SACT\_number\_cycles | 60 | 60 | 100.0% |  |  |
| HN\_3rd\_line\_pall\_SACT\_regime | 60 | 60 | 100.0% |  |  |
| HN\_3rd\_line\_pall\_SACT\_start\_date | 60 | 60 | 100.0% |  |  |
| HN\_3rd\_line\_pall\_SACT\_number\_cycles | 60 | 60 | 100.0% |  |  |
| HN\_4th\_line\_pall\_SACT\_regime | 60 | 60 | 100.0% |  |  |
| HN\_4th\_line\_pall\_SACT\_start\_date | 60 | 60 | 100.0% |  |  |
| HN\_4th\_line\_pall\_SACT\_number\_cycles | 60 | 60 | 100.0% |  |  |
| HN\_5th\_line\_pall\_SACT\_regime | 60 | 60 | 100.0% |  |  |
| HN\_5th\_line\_pall\_SACT\_start\_date | 60 | 60 | 100.0% |  |  |
| HN\_5th\_line\_pall\_SACT\_number\_cycles | 60 | 60 | 100.0% |  |  |

Table 5. Accuracy estimation results for chemotherapy data categories.

RT data was estimated to be more than 95% accurate for all categories. Of all radiotherapy data, 2 patients were found to have any incorrect data (Table 6), 1 due to an early termination of treatment not picked up by the structured radiotherapy data pull (as patient treated on the newer RT system (Aria)). For all 4 binary categories, PPV was 1, with NPV of 1 for all categories except RT replans and early termination (0.982 for both).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **RT data category** | **Assessed** | **Correct** | **% correct** | **PPV** | **NPV** |
| HNCRT | 60 | 60 | 100.0% | 1 | 1 |
| HNCRT\_course\_1\_site | 60 | 60 | 100.0% | 1 | 1 |
| HNC\_RT\_course\_1\_intent | 60 | 60 | 100.0% |  |  |
| HNC\_RT\_course\_1\_dose\_per\_Fr | 60 | 59 | 98.3% |  |  |
| HNC\_RT\_course\_1\_recorded\_total\_dose | 60 | 59 | 98.3% |  |  |
| HNC\_RT\_course\_1\_recorded\_fractions | 60 | 58 | 96.7% |  |  |
| HNC\_RT\_course\_1\_initially\_intended\_total\_dose | 60 | 59 | 98.3% |  |  |
| HNC\_RT\_course\_1 \_initially\_intended\_fractions | 60 | 59 | 98.3% |  |  |
| HNC\_RT\_course\_1\_start\_date | 60 | 60 | 100.0% |  |  |
| HNC\_RT\_course\_1\_end\_date | 60 | 60 | 100.0% |  |  |
| HNC\_RT\_course\_1\_elapsed\_time | 60 | 60 | 100.0% |  |  |
| HNC\_RT\_course\_1\_replans | 60 | 59 | 98.3% | 1 | 0.982 |
| HNC\_RT\_course\_1\_early\_termination | 60 | 59 | 98.3% | 1 | 0.982 |
| HNC\_RT\_course\_1\_early\_termination\_reason | 60 | 59 | 98.3% |  |  |
| HNCRT\_course\_2\_site | 60 | 60 | 100.0% |  |  |
| HNC\_RT\_course\_2\_intent | 60 | 60 | 100.0% |  |  |
| HNC\_RT\_course\_2\_dose\_per\_Fr | 60 | 60 | 100.0% |  |  |
| HNC\_RT\_course\_2\_total\_dose | 60 | 60 | 100.0% |  |  |
| HNC\_RT\_course\_2\_fractions | 60 | 60 | 100.0% |  |  |
| NonHNC\_RT | 60 | 60 | 100.0% |  |  |

Table 6. Accuracy estimation results for RT data categories.

Surgical data was estimated to be at least 95% accurate for all categories (Table 7). PPV and NPV was 1 for most surgical data categories, and more than 0.95 for all categories.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Surgical data category** | **Number assessed** | **Number correct** | **% correct** | **PPV** | **NPV** |
| Laryngectomy | 60 | 59 | 98.3% | 1.000 | 0.982 |
| Laryngectomy\_ProcedureDSC | 60 | 59 | 98.3% |  |  |
| Laryngectomy\_ProcedureDTS | 60 | 59 | 98.3% |  |  |
| Neck\_dissection | 60 | 57 | 95.0% | 0.952 | 0.949 |
| Neck\_dissection\_ProcedureDSC | 60 | 57 | 95.0% |  |  |
| Neck\_dissection\_ProcedureDTS | 60 | 57 | 95.0% |  |  |
| Tracheostomy | 60 | 59 | 98.3% | 1.000 | 0.978 |
| Tracheostomy\_Procedure\_description | 60 | 59 | 98.3% |  |  |
| Tracheostomy\_First\_Procedure\_Date | 60 | 59 | 98.3% |  |  |
| Mandible\_excision | 60 | 59 | 98.3% | 1.000 | 0.982 |
| Mandible\_excision\_ProcedureDSC | 60 | 59 | 98.3% |  |  |
| Mandible\_excision\_ProcedureDTS | 60 | 59 | 98.3% |  |  |
| Maxillectomy | 60 | 60 | 100.0% | 1.000 | 1.000 |
| Maxillectomy\_ProcedureDSC | 60 | 60 | 100.0% |  |  |
| Maxillectomy\_ProcedureDTS | 60 | 60 | 100.0% |  |  |
| Total\_Glossectomy | 60 | 60 | 100.0% | 1.000 | 1.000 |
| Total\_Glossectomy\_ProcedureDSC | 60 | 60 | 100.0% |  |  |
| Total\_Glossectomy\_ProcedureDTS | 60 | 60 | 100.0% |  |  |
| Partial\_Glossectomy | 60 | 60 | 100.0% | 1.000 | 1.000 |
| Partial\_Glossectomy\_ProcedureDSC | 60 | 60 | 100.0% |  |  |
| Partial\_Glossectomy\_ProcedureDTS | 60 | 60 | 100.0% |  |  |
| Oral\_cavity\_surgery | 60 | 60 | 100.0% | 1.000 | 1.000 |
| Oral\_cavity\_surgery\_ProcedureDSC | 60 | 60 | 100.0% |  |  |
| Oral\_cavity\_surgery\_ProcedureDTS | 60 | 60 | 100.0% |  |  |
| Tonsillectomy | 60 | 60 | 100.0% | 1.000 | 0.982 |
| Tonsillectomy\_ProcedureDSC | 60 | 60 | 100.0% |  |  |
| Tonsillectomy\_ProcedureDTS | 60 | 60 | 100.0% |  |  |
| Pharyngectomy | 60 | 60 | 100.0% | 1.000 | 1.000 |
| Pharyngectomy\_ProcedureDSC | 60 | 60 | 100.0% |  |  |
| Pharyngectomy\_ProcedureDTS | 60 | 60 | 100.0% |  |  |
| Salivary\_gland\_surgery | 60 | 59 | 98.3% | 1.000 | 0.982 |
| Salivary\_gland\_surgery\_procedure\_description | 60 | 59 | 98.3% |  |  |
| Salivary\_gland\_surgery\_First\_procedure\_date | 60 | 59 | 98.3% |  |  |
| Nasal\_cavity\_sinus\_surgery | 60 | 59 | 98.3% | 1.000 | 1.000 |
| Nasal\_cavity\_sinus\_surgery\_ProcedureDSC | 60 | 60 | 100.0% |  |  |
| Nasal\_cavity\_sinus\_surgery\_ProcedureDTS | 60 | 60 | 100.0% |  |  |
| Orbital\_Exenteration | 60 | 60 | 100.0% | 1.000 | 1.000 |
| Orbital\_Exenteration\_Procedure\_Description | 60 | 60 | 100.0% |  |  |
| Orbital\_Exenteration\_ProcedureDTS | 60 | 60 | 100.0% |  |  |
| Salvage\_procedure\_CogStack | 50 | 49 | 98.0% | 1.000 | 0.981 |
| Electrochemotherapy | 60 | 60 | 100.0% |  |  |
| Electrochemotherapy\_procedure\_description | 60 | 60 | 100.0% |  |  |
| Electrochemotherapy\_first\_procedure\_date | 60 | 60 | 100.0% |  |  |
| Gastrostomy | 60 | 58 | 96.7% | 0.955 | 0.974 |
| Gastrostomy\_procedure\_description | 60 | 58 | 96.7% |  |  |
| Gastrostomy\_date | 60 | 58 | 96.7% |  |  |
| Gastrostomy\_removal | 14 | 14 | 100.0% | 1.000 | 1.000 |
| Gastrostomy\_removal\_ProcedureDSC | 14 | 14 | 100.0% |  |  |
| Gastrostomy\_removal\_First\_Procedure\_Date | 14 | 14 | 100.0% |  |  |

Table 7. Accuracy estimation results for Surgical data categories.

Outcome data was estimated to be 100% accurate for all categories except two (last follow up date, treatment failure status, Table 8). One patient was found to have a later follow up than had been picked up on the structured data pull and 1 patient was found to have had a new lung cancer rather than a HNC treatment failure. These were the only errors seen in all data assessed. PPV and NPV was 1 for binary data (alive/deceased status) for follow up.

|  |  |  |  |
| --- | --- | --- | --- |
| **Outcomes data point** | **Number assessed** | **Number correct** | **% correct** |
| LastFUdate | 60 | 59 | 98.3% |
| Deceased | 60 | 60 | 100.0% |
| DeathDTS | 60 | 60 | 100.0% |
| Cause\_of\_death | 3 | 3 | 100% |
| Cause\_of\_death\_other | 3 | 3 | 100% |
| CogStack\_CR\_and\_radical\_RT | 9 | 9 | 100% |
| 3m\_RT\_response | 30 | 30 | 100% |
| 3m\_imaging\_modality | 18 | 18 | 100% |
| 3m\_imaging\_date | 24 | 24 | 100% |
| 6m\_RT\_response | 27 | 27 | 100% |
| 6m\_imaging\_modality | 20 | 20 | 100% |
| 6m\_imaging\_date | 22 | 22 | 100% |
| Biopsy\_post\_RT | 0 | 0 | n/a |
| Bx\_date | 0 | 0 | n/a |
| Salvage\_neck\_dissection | 13 | 13 | 100% |
| ND\_date | 13 | 13 | 100% |
| Failure | 18 | 17 | 94.4% |
| Failure\_date | 15 | 15 | 100% |
| Failure\_site\_summary | 15 | 15 | 100% |
| Primary\_failure\_site\_details | 13 | 13 | 100% |
| Primary\_recurrence\_date | 13 | 13 | 100% |
| Primary\_recurrence\_Rx | 13 | 13 | 100% |
| Primary\_recurrence\_intent | 13 | 13 | 100% |
| Nodal\_or\_locoregional\_failure\_site | 14 | 14 | 100% |
| Nodal\_or\_locoregional\_non\_primary\_recurrence\_date | 14 | 14 | 100% |
| Nodal\_recurrence\_Rx | 13 | 13 | 100% |
| Nodal\_recurrence\_intent | 13 | 13 | 100% |
| Metastatic\_site | 13 | 13 | 100% |
| Met\_recurrence\_date | 14 | 14 | 100% |
| Met\_recurrence\_Rx | 14 | 14 | 100% |
| Met\_recurrence\_intent | 14 | 14 | 100% |
| Recurrence\_after\_RT | 19 | 19 | 100% |
| Time\_to\_failure\_after\_RT | 19 | 19 | 100% |
| In\_field\_recurrence\_\_high\_dose | 17 | 17 | 100% |
| In\_field\_recurrence\_\_ppx\_dose | 17 | 17 | 100% |

Table 8. Accuracy estimation results for outcome data categories.

For data categories that were direct structured EHR data pulls that did not undergo transformation, a sample of 20 for each was assessed (MDM data, bloods data, CTCAE, PROMS data, RT OAR/target volume dose parameter data). All data was 100% accurate corresponding to the EHR source data, ensuring no unexpected data alterations had inadvertently occurred following initial extraction from source EHR.

*Data completeness assessment*

Full data completeness results can be seen in Appendix 5. Data completeness varied according to data availability across different data categories, however across the key categories of diagnosis and oncology first visit dates, disease site, histopathology, initial treatment received, treatment data (surgery, RT and chemotherapy), and follow up (last follow up and death status and death) data was 100% complete.

*Data consistency assessment*

Data was reviewed for consistency during the evaluation process to ensure no inconsistencies seen. Terminology within the clinical dataset was consistent across all data categories according to the rules set up before dataset build. This included all data known to not exist (such as a second course of RT for a patient who only ever had one course of RT) being labelled as ‘Not applicable’, and all data with missing values labelled as ‘NULL’. During the final validation process, a small number of inconsistencies within the chemotherapy data were found. Examples included:

* Slight discrepancy for induction chemotherapy regimen labelled (2 patients '…then…' rather than ‘+’)
* NACT: GA201 anti-EGFR antibody (trial) also referred to as 'GA201 cetuximab trial for one patient

The above inconsistencies were all corrected. No other inconsistencies were found within in the dataset during the validation and verification process.

**Discussion**

We built a large clinical dataset containing over 1.7 million data points of demographic, co-morbidity, diagnosis, treatment, laboratory and outcome data for our complete cohort of the 2895 HNC patients meeting our inclusion/exclusion criteria (Figure 6). Given the completeness of our population by cross-referencing different data sources to corroborate the final patient population, we believe this dataset is a relevant longitudinal dataset representing a diverse real world UK HNC oncology patient population treated at a large NHS oncology centre over more than a decade. Both during the dataset build and validation phases, care was taken to respect the data quality dimensions set out by regulatory bodies including NICE (17) to ensure reliable and relevant data of robust quality to enable data use for meaningful RWE generation. By undertaking the dataset validation and verification checks, we have demonstrated the accuracy, completeness and consistency of our clinical dataset across critical variables.



Figure 6. Summary of data present in retrospective HNC clinical dataset (in the style of mCODE figure (Figure 1).

**Strengths of dataset: Data curated**

Our dataset contains a number of strengths in terms of data categories present that other reported large HNC RWD datasets have not contained (27,28). Our dataset is comprehensive across key data subcategories of demographics, co-morbidities, diagnosis factors, treatments received, laboratory results and outcomes, providing a comprehensive patient-factor focused clinical dataset and therefore facilitating robust RWE generation allowing consideration of confounding factors. As opposed to limited datasets designed for a solitary research question, this is a scaled dataset with potential application to many different research questions. Of particular note our dataset contains over 300,000 data points for CTCAE and RTOG toxicity data documented prospectively at the point of care by Clinical Oncologists over 13 years. Following literature review, to our knowledge this represents one of the largest amalgamations of CTCAE and RTOG toxicity over a prolonged time-period for a HNC dataset. Further, the inclusion of over 60000 data points of PROMs data is a further novel strength of our dataset providing a range of potential research opportunities that can be explored with this data. For example, there has been limited research evaluating how HNC CTCAE/RTOG toxicity correlates with PROMs. This showcases just one example that our RWD may have in complementing clinical trial data within the HNC research setting. A further strength of our dataset is that we have linked relevant non-EHR UK deprivation data to patient post code data allowing further consideration of research questions investigating how deprivation impacts on the HNC population, which remains an unmet need (29).

One of the biggest novel strengths of our clinical dataset is that it will be linked to diagnostic imaging, and RT imaging and treatment data, facilitating multi-modal research projects. While other groups outside the UK have developed multi-modal datasets of similar size (in terms of patient numbers), such as the RADCURE dataset in Canada (30), our dataset contains a much broader range of clinical data categories.

**Strengths of dataset: Dataset build methodology**

The clinical dataset build methodology contains several strengths to ensure a complete, consistent and accurate dataset has been built that can be used to generate robust RWE. Firstly, by cross-referencing different data extracts from different sources, allowing targeted and feasible verification checks (for example checking when there is limited disagreement of data for the same data category from different sources such as for surgical procedure data), this allows assembly of the most comprehensive and accurate data possible. For example, chemotherapy data was taken principally from a structured data extract however this was also cross-checked with CogStack’s output for concepts such as ‘induction chemotherapy’ and ‘concomitant chemotherapy’ run for our entire patient population. This process facilitated the identification of a minority of missing patients and subsequent curation of any missing chemotherapy data to optimise our data quality.

The use of CogStack’s NLP model was a further innovative approach used to build this dataset. Based on findings from previous evaluation work, CogStack was used to curate data for validated SNOMED CT concepts. To our knowledge this is the first NHS oncology clinical dataset part-curated using NLP outside of a commercial company setting. This is a method used by leading commercial companies such as Flatiron Health (3), a company that was sold to Roche in 2018 for $1.8 billion (30). Despite having a fraction of the resources that would presumably be available to a multi-billion dollar company with a revenue of over $200 million per year, we have been able to apply similar approaches within the context of an NHS department. We hope that this approach demonstrates that NLP can be used in conjunction with structured data extracts within the NHS for clinical data curation.

Finally, this dataset build was undertaken with close attention to key data quality dimensions set out within regulatory body guidance on the use of RWD for RWE to ensure our data curation process and evaluation was undertaken following gold standard approaches followed by the RWD industry. As demonstrated by high accuracy, PPV and NPV figures (more than 98% for the vast majority of data categories), and completeness of 100% for the most critical variables, we have built a robust clinical dataset using multiple approaches. We believe that our dataset provides an example to other NHS oncology centres that robust large scale clinical datasets can be built in-house without the necessity for agreements with commercial partners which may lead to loss of rights over data usage.

**Limitations of dataset: Data curated**

There are a number of areas within our clinical dataset that currently demonstrate deficiencies which we hope to rectify in the future. This data has not yet been curated for our dataset because this was not available within a structured data extract, nor did CogStack demonstrate the capability to cover these missing areas. Firstly, although we currently have complete and accurate last follow up and death data, as well as toxicity outcome form data, we lack completeness of other disease outcome data that would have potential importance for future research projects. These disease outcome data points include response to treatment at different time points, for example, 3 months after radiotherapy, site of disease recurrence/relapse, or cause of death data. Outstanding data will likely require manual curation, however a significant amount of data was curated using existing GT dataset and HN assessment form data in conjunction with abstraction of data from pre-existing data within the cohort (for example we know those who had radical treatment then had palliative SACT must have relapsed). A similar area of weakness that will require some further manual data curation was histopathological data categories such as presence of lymphovascular invasion, extracapsular extension, depth of invasion and surgical margin status, which again were not available from a structured data extract or CogStack.

Other challenges encountered during the dataset build included the presence of duplicates within the datasets, where patients suffered from multiple HNC (70 patients). This has the potential to add a level of confusion, particularly as it was sometimes unclear even from the medical notes whether the second cancer was a new primary or a recurrent tumour. This issue was limited by following standardised rules for dataset curation as previously set out.

Further difficulties were seen for more historic patients (particularly 2010-2012 patients), where data completion and accuracy was observed during the data curation process to be worse than more recent data. There are a number of reasons that likely explain the limitations of more historic patient data within our dataset, most prominently the less consistent and more infrequent use of EHRs at this point in time (for example, chemotherapy prescriptions were not completely electronic until 2012). This observation does however demonstrate the potential benefits that EHRs have brought to accurate large scale RWD curation.

A further challenge we encountered during dataset curation was where some patients had had oncological treatment (surgery, RT or chemotherapy) at another centre. Details of this treatment therefore did not appear on structured data extracts from EHRs. To counter this issue, a number of strategies were employed to ensure as much data was present for treatments elsewhere as possible. Firstly, CogStack output for different treatment modalities was reviewed and could be cross-referenced with structured data extracts from our own EHRs. Secondly, by indirect benchmarking (for example, a patient was shown to have had a post-operative radiotherapy course but no surgery) any gaps in data where patients had treatment elsewhere could be further rectified. Finally, during data validation both during and after the data curation process if a patient was found to have had treatment elsewhere, this was manually annotated within the clinical dataset.

**Limitations of dataset: Dataset build methodology**

Whilst we believe we undertook the optimum dataset build based on our capabilities, resources, and data availability there are a number of limitations to our methodology. Firstly, whilst estimations of accuracy, PPV and NPV were obtained for data categories assessed, these assessments were based on relatively small samples (60 of 2895 patients) and were not performed for all 1800 categories within the dataset. Due to the dataset size and time available, it was not feasible to carry out a more extensive data validation and verification process. This is an issue that has been highlighted with RWD validation processes even by large commercial companies with much larger resources, with prioritisation of critical variables required (3). A further counterpoint to this limitation was that although only 60 data points for a data category may have been checked on the final dataset, data was also constantly checked during the data curation process to ensure accuracy.

Although CogStack was used during the dataset build, the majority of data within the final dataset was curated using other methods. This was because CogStack’s performance based on our evaluation was good enough for 50 SNOMED CT concepts, and for a number of these ultimately a structured data extract with more detail was found from EHR data tables on the EDW. For example, whilst CogStack’s performance was felt to be good for several different surgical procedures, this data was ultimately found on an EPR data table on the EDW, which contained more information (date of procedure, exact procedure name performed) than CogStack’s output which was a simple binary present or absent denotation. The experience and limited use of CogStack for our dataset build reflects limitations established during our team’s previous CogStack evaluation work. Ultimately CogStack was directly used for population of co-morbidity data categories (felt not to be reliably coded on the EPR data table on the EDW) as well as histopathological concepts not available from structured data pulls such as perineural invasion. However, despite this limited direct use, CogStack was also indirectly an invaluable tool for the dataset curation. As discussed, CogStack output was compared with structured data extracts and differences seen checked and verified or rectified accordingly to ensure data accuracy and completeness was optimised.

The final limitation of our dataset relates to data imputation at source. Our dataset relies on the principle that source data entered by healthcare professionals is always accurate when clearly data may sometime be entered erroneously or not entered at all, an issue that can affect both structured and unstructured data formats. For example, a patient may be coded to have a tongue cancer (oral cavity) when in reality they had a base of tongue cancer (oropharynx), which represents a different disease process necessitating different treatment. A further example would be that a patient never had a 1 year toxicity assessment form completed when they attended for their 1 year post-treatment follow up appointment and this data was therefore missing. These are intrinsic limitations of using RWD as described above, that hopefully will improve with time. To counter this issue we corroborated as many different data sources for each data category to ensure each patient’s data was as complete and accurate as feasible.

**Future work**

*Retrospective dataset*

Our clinical data for each patient is linkable with corresponding patient diagnostic and radiotherapy imaging data on XNAT for future research projects. By applying recorded and reproducible filters to ‘search’ our dataset for patients required for a specific use case (for example, all salivary gland cancer patients treated with radiotherapy), anonymised selected data collections are available for approved requests through a data access pipeline safeguarded by our centre’s XNAT Research Access Panel (XRAP). These will be supplied to users in the form of a CSV file with anonymised subject IDs corresponding to anonymised patient imaging data supplied via XNAT. To improve usability, a short ‘User Guide Summary’ document has been developed for clinical data users to complement the more extensive data dictionary.

Whilst we have built a clinical dataset that we believe to be relevant and reliable enough for RWE, we hope to continue to expand and improve this dataset to address outstanding or limited areas as described above. Firstly, key outstanding critical variables including cause of death, outcome data categories and missing histopathological factors will be manually curated in a targeted process. Beyond these key areas, other desirable areas of data curation have been identified that can be improved over time. As the dataset grows, further completeness assessments will be undertaken every 6 months to guide future areas of requiring focus for improvement. We also hope to link our clinical and imaging dataset to biobank tissue data for further data aspects including genomics. Furthermore, future data for this patient cohort related to toxicity (CTCAE/RTOG and PROMs) data, as well as further treatment and outcomes data will be updated with time once data pipelines are available from the new EHR (Epic) system.

*Prospective dataset*

Beyond improving and updating our retrospective clinical dataset, we plan to build a prospective dataset with capability for automatic data pipeline updates extracting relevant structured data points for future patients under the HNC oncology team meeting the same inclusion and exclusion criteria as our retrospective cohort. Similar methodology will be employed as was used for the retrospective clinical dataset build, however this will need to be adapted accordingly based on the data available from Epic and the transformation any of this data requires following structured data extraction. The practicalities of undertaking this work will be clearer once the data warehouse containing data tables from Epic (the Guy’s Data Warehouse (GDW)) is fully functional and optimised following the ongoing initial phases of the Epic EHR rollout. Work is ongoing to ensure the same data is captured for our prospective patients as our retrospective patients to facilitate ongoing RWD collection and therefore RWE generation.

**Summary**

We have built a large clinical dataset of demographic, co-morbidity, diagnosis, treatment, laboratory and outcome data for our complete retrospective HNC oncology cohort using a combination of data curation strategies including EHR structured data extracts and NLP. This clinical dataset underwent a validation and verification process, in accordance with NICE RWE generation framework guidelines, showing excellent accuracy, completeness and consistency of critical variables. Work is ongoing to improve this clinical dataset further to complete any missing data categories of relevance as well ensure this dataset is updated using automated data pipelines from the new EHR system (Epic) in the future. Further work is ongoing to build a prospective data flow to ensure this dataset is added to with longitudinal data from future HNC oncology patients. We believe this clinical dataset can underpin many future research projects investigating different aspects of HNC to facilitate robust RWE generation.

**References**

1. Sdeb S, Lee A, van Dijk LV, Frank S, Fuller CD, Blanchard P. NTCP Modeling of Late Effects

for Head and Neck Cancer: A Systematic Review. Int J Part Ther. 2021 Jun 25;8(1):95-107.

doi: 10.14338/20-00092. PMID: 34285939; PMCID: PMC8270107.

2. US Food & Drug Administration. *Real-World Evidence.* <https://www.fda.gov/science-research/real-world-evidence/center-biologics-evaluation-and-research-center-drug-evaluation-and-research-real-world-evidence#:~:text=As%20defined%20by%20FDA%2C%20real,derived%20from%20analysis%20of%20RWD>. [Accessed 10 May 2024].

3. Castellanos EH, Wittmershaus BK, Chandwani S. Raising the Bar for Real-World Data in Oncology: Approaches to Quality Across Multiple Dimensions. JCO Clin Cancer Inform. 2024 Jan;8:e2300046. doi: 10.1200/CCI.23.00046. PMID: 38241599; PMCID: PMC10807898.

4. Cancer Research UK. (2022). *Our research data strategy.* hnps://www.cancerresearchuk.org/funding-for-researchers/research-opportunides-in- data-science [Accessed 5th November 2023].

5. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, LaVange L, Marinac-Dabic D, Marks PW, Robb MA, Shuren J, Temple R, Woodcock J, Yue LQ, Califf RM. Real-World Evidence - What Is It and What Can It Tell Us? N Engl J Med. 2016 Dec 8;375(23):2293-2297. doi: 10.1056/NEJMsb1609216. PMID: 27959688.

6. National Cancer Institute. *Surveillance, Epidemiology and End Results (SEER).* <https://seer.cancer.gov/about/factsheets/SEER_Overview.pdf> [Accessed on 10 May 2024].

1. NHS England. *National Cancer Registration and Analysis Service (NCRAS).* <https://digital.nhs.uk/ndrs/about/ncras> [Accessed on 9 May 2024].
2. Edwards L, Pickett J, Ashcroft DM, Dambha-Miller H, Majeed A, Mallen C, Petersen I, Qureshi N, van Staa T, Abel G, Carvalho C, Denholm R, Kontopantelis E, Macaulay A, Macleod J. UK research data resources based on primary care electronic health records: review and summary for potential users. BJGP Open. 2023 Sep 19;7(3):BJGPO.2023.0057. doi: 10.3399/BJGPO.2023.0057. PMID: 37429634; PMCID: PMC10646196.
3. Makady A, de Boer A, Hillege H, Klungel O, Goettsch W, et al. What is

real-world data? A review of definitions based on literature and stake-

holder interviews. Value Health. 2017;20(7):858–65

1. Osterman TJ, Terry M, Miller RS. Improving Cancer Data Interoperability: The Promise of the Minimal Common Oncology Data Elements (mCODE) Initiative. JCO Clin Cancer Inform. 2020 Oct;4:993-1001. doi: 10.1200/CCI.20.00059. PMID: 33136433; PMCID: PMC7713551.
2. HL7 International. *Minimal Common Oncology Data Elements (mCODE) Implementation Guide Version 4.* <https://build.fhir.org/ig/HL7/fhir-mCODE-ig/> [Accessed on 11 May 2024].
3. Ross MK, Wei W, Ohno-Machado L. "Big data" and the electronic health record. Yearb Med Inform. 2014 Aug 15;9(1):97-104. doi:10.15265/IY2014-0003. PMID: 25123728; PMCID: PMC4287068.
4. HealthCatalyst. (2023). *HealthCatalyst Data Platform*. https://www.healthcatalyst.com/offerings/platform [Accessed 5th November 2023].
5. Rahman P, Fabbri D. Semi-Automated Data Curadon from Biomedical Literature. AMIA Annu Symp Proc. 2023 Apr 29;2022:884-891. PMID: 37128469; PMCID: PMC10148326.
6. Wang L, Fu S, Wen A, Ruan X, He H, Liu S, Moon S, Mai M, Riaz IB, Wang N, Yang P, Xu H, Warner JL, Liu H. Assessment of Electronic Health Record for Cancer Research and Patient Care Through a Scoping Review of Cancer Natural Language Processing. JCO Clin Cancer Inform. 2022 Jul;6:e2200006. doi: 10.1200/CCI.22.00006. PMID: 35917480; PMCID: PMC9470142.
7. Hong JC, Fairchild AT, Tanksley JP, Palta M, Tenenbaum JD. Natural language processing for abstraction of cancer treatment toxicities: accuracy versus human experts. JAMIA Open. 2020 Dec 5;3(4):513-517.doi: 10.1093/jamiaopen/ooaa064. PMID: 33623888; PMCID: PMC7886534.
8. National Institure of Clinical Excellence. NICE real-world evidence framework: Assessing data suitability. 2022. https://www.nice.org.uk/corporate/ecd9/chapter/assessing-data-suitability#assessing-data-suitability [Accessed 9 May 2024].
9. US Food & Drug Administration. *Considerations for the Use of Real-World Data and Real-World Evidence To Support Regulatory Decision-Making for Drug and Biological Products. 2023.* <https://www.govinfo.gov/content/pkg/FR-2023-12-19/pdf/2023-27852.pdf> [Access on 9 May 2024].
10. European Medicines Agency. European Medicines Agency: Data Quality Framework for EU Medicines Regulation. 2022.https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/data-quality-framework-eu- medicines-regulation\_en.pdf [Accessed on 11 May 2024].
11. Herrick R, Horton W, Olsen T, McKay M, Archie KA, Marcus DS. XNAT Central: Open sourcing imaging research data. Neuroimage. 2016 Jan 1;124(Pt B):1093-1096. doi: 10.1016/j.neuroimage.2015.06.076. Epub 2015 Jul 2. PMID: 26143202; PMCID: PMC4965359.Denney MJ, Long DM, Armistead MG, Anderson JL, Conway BN. Validating the extract, transform, load process used to populate a large clinical research database. Int J Med Inform. 2016 Oct;94:271-4. doi: 10.1016/j.ijmedinf.2016.07.009. Epub 2016 Jul 29. PMID: 27506144; PMCID: PMC5556907.
12. Sambasivan K, Sassoon I, Thavaraj S, Kennedy R, Doss G, Michaelidou A, Odell E, Sandison A, Hall G, Morgan P, Collins LHC, Lyons A, Cascarini L, Fry A, Oakley R, Simo R, Jeannon JP, Petkar I, Reis Ferreira M, Kong A, Lei M, Guerrero Urbano T. TNM 8 staging is a better prognosticator than TNM 7 for patients with locally advanced oral cavity squamous cell carcinoma treated with surgery and post-operative radiotherapy. Radiother Oncol. 2021 Jul;160:54-60. doi: 10.1016/j.radonc.2021.04.003. Epub 2021 May 3. PMID: 33845044.
13. De Felice F, Bird T, Michaelidou A, Jeannon JP, Simo R, Oakley R, Lyons A, Fry A, Cascarini L, Asit A, Thavaraj S, Reis Ferreira M, Petkar I, Kong A, Lei M, Guerrero Urbano T. Clinical outcomes in relapsed oropharyngeal cancer after definitive (chemo) radiotherapy. Oral Dis. 2023 Mar;29(2):595-603. doi: 10.1111/odi.13985. Epub 2021 Aug 27. PMID: 34338394.
14. National Health Service. *NHS Data Model and Dictionary.* 2024. <https://www.datadictionary.nhs.uk/data_elements/date_of_primary_cancer_diagnosis__clinically_agreed_.html#:~:text=This%20will%20normally%20be%20the,of%20the%20Multidisciplinary%20Team%20Meeting> [9 May 2024].
15. Ministry of Housing, Communities & Local Government. *English indices of deprivation 2019.* <https://imd-by-postcode.opendatacommunities.org/imd/2019> [9 May 2024].
16. Moss C, Haire A, Cahill F, Enting D, Hughes S, Smith D, Sawyer E, Davies A, Zylstra J, Haire K, Rigg A, Van Hemelrijck M. Guy's cancer cohort - real world evidence for cancer pathways. BMC Cancer. 2020 Mar 17;20(1):187. doi: 10.1186/s12885-020-6667-0. PMID: 32178645; PMCID: PMC7077127.
17. National Cancer Institute. *Head & Neck Datasets*. <https://cdas.cancer.gov/datasets/plco/15/> [Accessed on 9 May 2024].
18. [Grossberg AJ, Mohamed ASR, Elhalawani H, Bennett WC, Smith KE, Nolan TS, Williams B, Chamchod S, Heukelom J, Kantor ME, Browne T, Hutcheson KA, Gunn GB, Garden AS, Morrison WH, Frank SJ, Rosenthal DI, Freymann JB, Fuller CD. Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy. Sci Data. 2018 Sep 4;5:180173. doi: 10.1038/sdata.2018.173. Erratum in: Sci Data. 2018 Nov 27;5(1):1. PMID: 30179230; PMCID: PMC6190723.](https://www.cancerimagingarchive.net/collection/hnscc/)
19. Unger JM, Moseley AB, Cheung CK, Osarogiagbon RU, Symington B, Ramsey SD, Hershman DL. Persistent Disparity: Socioeconomic Deprivation and Cancer Outcomes in Patients Treated in Clinical Trials. J Clin Oncol. 2021 Apr 20;39(12):1339-1348. doi: 10.1200/JCO.20.02602. Epub 2021 Mar 17. PMID: 33729825; PMCID: PMC8078474.
20. Flatiron Health. *Roche Completes Acquisition of Flatiron Health.* 2018. <https://resources.flatiron.com/press/press-release/roche-completes-acquisition-of-flatiron-health> [Accessed on 9 May 2024].
21. Chen M, Wang K, Wang J. Advancing Head and Neck Cancer Survival Prediction via Multi-Label Learning and Deep Model Interpretation. ArXiv [Preprint]. 2024 May 9:arXiv:2405.05488v1. PMID: 38764586; PMCID: PMC11100915.