SOP: Transfer of Radiotherapy Data into RTDLHN

**Purpose:** Document the construction of the radiotherapy section of the Radiotherapy Head and Neck Data Lake (RTDLHN)

**Scope:** Inclusion/Exclusion criteria, processes to gather data, database structure, quality control and checking, known deficits, scope for future improvement.

# Abstract

Purpose/Objective   
The optimal performance of machine learning (ML) models and their generalisability relies on the quality of the data for model construction. Retrospective and prospective collection of high-quality data for research use whilst respecting data protection restrictions and patient privacy remains a challenge in the clinical environment. Currently, months of laborious extraction and clinical annotation are often necessary before data analysis can begin to ensure the completeness, accuracy, and usefulness of data sets for ML. We present a novel project architecture, utilising open-source software, to facilitate a fast and efficient production of ML models from an institutional federated data lake containing high quality Head and Neck Cancer (HNC) imaging and Radiotherapy (RT) data with relevant clinical annotations. The data lake and data access pipeline will dramatically reduce the time associated with the production of ML models and Real-World Evidence (RWE) reporting. The aim of this article is to provide a meaningful framework on how to set up a database to gather and use patient data in line with the best information governance principles.

Materials and Methods/Construction and Content

XNAT is a powerful open-source platform capable of storing and managing medical imaging and associated data. It provides import, archiving, processing, search and secure distribution facilities. Within Guy’s and St Thomas’ NHS Foundation Trust (GSTT), it forms a part of the local secure enclave for the purposes of federated learning in artificial intelligence projects.

## Results and Recommendations

We have created a secure XNAT data lake of consenting HNC RT patients hosted by the Clinical Scientific Computing (CSC) team at our organisation. We provide a list of recommendations and considerations for other institutes or cancer sites to follow which include careful consideration of the infrastructure and longevity of support of the technical solution, patient and data selection criteria, and emphasis on the preservation of the integrity of patient consent and patients’ rights to privacy and data protection.

Conclusions   
We have created a secure and extensible imaging and HNC RT cancer database. This database set-up promises to be an exceedingly useful tool for research, revolutionising the time and cost associated with the production of machine learning models, making the process safer, faster and more efficient.

# Background

Cancer remains one of the leading causes of death in the UK and it is estimated that 1 in 2 people will be diagnosed with cancer at some point during their lifetime [1]. With over 1000 new diagnoses a day, the burden of disease is large and, whilst 50% patients will survive cancer for 10 or more years, the side effects from treatment can result in significant mortality and morbidity, in addition to cost to healthcare systems.

Radiotherapy (RT) is a key treatment of HNC, delivered with curative intent in a large number of patients but with poor outcomes in a significant proportion who either recur and succumb to their disease or, despite advances in RT imaging, planning and treatment delivery, have significant permanent long-term toxicities. Predictive models integrating outcome and RT data underpin many of the dose-volume constraints used in clinical practice to limit doses to organs at risk (Normal Tissue Complication Probability Modelling- NTCP). Models have also been developed to try and predict tumour control outcomes (Tumour Control Probability Models). These models, however, vary in their level of complexity, many lack validation and most are not used in the clinic. Recently, machine learning (ML) has been proposed as a way to improve on previous models. This is because it provides the potential to model multifactorial RT outcomes, can incorporate RT imaging (i.e. dose maps) in addition to the DVH and fractionation data used by traditional modelling, is able to consider multiple clinical and biological prognostic factors and can use large clinical datasets. The resulting improved predictive models could enable individualized RT with the aim to increase tumour control whilst reducing radiation toxicity and improve patients’ quality of life.

Randomized Clinical Trials (RCTs) provide the highest level of evidence to define the efficacy of treatments. However, stringent eligibility criteria including age, performance status and/or absence of clinically relevant comorbidities create selection bias and it has been shown that real world population outcomes do not match those of RCTs [5]. In this context, analysis of real-world data (RWE) can help answer relevant clinical and policy questions that cannot be directly or completely answered using data from RCTs [5] or where these are simply not available. RWE can describe treatment outcomes from a more heterogeneous population, in particular through the inclusion of under-represented groups, i.e. due to age or race, and/or those excluded i.e. due to certain co-morbidities. RWE can also provide data for a longer period of follow-up and enable the long-term evaluation of treatment approaches [6]. Alternative data collection methodologies include longitudinal observation studies, registry based clinical trials and prospective databases [7].

Predictive modelling using ML requires large quantities of accurate, discrete clinical data allied to RT structure structured in a research ready format nor annotated and ML model development necessitates months of laborious extraction and annotation [8]. We therefore aimed to establish data infrastructure to enable meaningful and rapid scientific advances whilst preserving the integrity of patient consent.

A database of clinically annotated diagnostic imaging and RT treatment data that can be readily accessible and usable under a transparent and safe system of governance and stewardship as well as patient consent and ethics approval has been constructed adhering to the FAIR guiding principles regarding the use and reuse of digital data [11]. This document describes our experience of setting up the retrospective cohort of the database and recommendations for technical and practical considerations to construct a valuable research tool.

# Material and Methods/Construction and Content

## Data lake infrastructure

Extensible NeuroImaging Archival Toolkit (XNAT) [12] is a powerful open-source platform capable of storing and managing medical imaging and associated clinical data. It provides import, archiving, processing and secure distribution facilities. Whilst XNAT was originally designed for neuroimaging DICOM data, developments have enabled the handling of RT data both nationally [13] and internationally to facilitate data sharing [14].

Within our organisation, XNAT forms a part of the local secure enclave for the purposes of federated learning in artificial intelligence projects within Clinical Scientific Computing. This neat solution is independent of electronic patient record provider and RT vendor to enable incorporation of legacy RT and clinical data as well as diagnostic imaging data.

A DICOM Q/R service was set up on Aria using the Database Service option from Varian. This enables the ingestion of RT images, structure sets, dose, RT plans, any additional imaging and spatial registrations used during treatment planning and, for patients treated through Aria, the on-set imaging (CBCTs) and RT Beams Treatment Record information. XNAT allows searching across each file type and custom forms to allow the location of data of interest for specific research use [11]. Within GSTT’s XNAT environment, we have created a specific Head and Neck XNAT project that holds identifiable patient data; this is necessary so that each patient dataset can be updated with new events. This project is referred to as the Radiotherapy Head and Neck Data Lake (RTDLHN).

## Patient Selection

### Retrospective cohort creation

HNC patients seen by the clinical oncology H&N team between the introduction of intensity modulated radiotherapy (IMRT) for radical patients (March 2011) and the introduction of a new hospital-wide electronic health record system (EPIC) (October 2023) were included within the retrospective cohort. Patients with skin primaries, paraganglioma, thyroid & parathyroid and any other non-HNC patients were excluded, in addition to patients who went on to have their treatment elsewhere and patients treated under different teams (e.g. lymphoma). Patients were assigned a single globally unique and persistent identifier [11], their NHS number, within the database to enable alignment between each data source and prevent the splitting of patient records into multiple records. Patients without NHS numbers (n=8) were given a separate unique identifier.

### Radiotherapy patient selection

Patients who received Radiotherapy at GSTT had their RT treatment records retrieved and stored. These can be found within .csv file attached to the XNAT data lake.

Over the time period of data collection (March 2011-October 2023) there were 3 treatment planning systems, 2 record and verify systems, 3 locations of on-set imaging storage and numerous incremental improvements to treatment planning techniques. In 2017 Eclipse was introduced as the only treatment planning system with Mosaiq as the corresponding record and verify system. Mosaiq was replaced with Aria in October 2021. Figure 1 shows the timeframe over which each system was operational.

***Figure 1: Legacy and current radiotherapy plan, treatment record and imaging stores with corresponding time periods labelled for HNC RT patients.***

In 2011, we treated patients using Monaco to create IMRT treatment plans, delivering the patient plan using Mosaiq as the record and verify system and imaging using XVI. 3D-conformal radiotherapy (3D-CRT) plans for palliative patients or simpler patient plans (e.g. larynx) were constructed in XiO and delivered through Mosaiq. From 2017 onwards, after moving HNC Radiotherapy to the Cancer Centre (May 2017), Eclipse was used to create all patient treatment plans, both IMRT/VMAT and conformal and we continued to deliver the plans using Mosaiq and the Mosaiq Data Director was used to store patient imaging. In October 2021 we continued to use Eclipse to create all patient treatment plans but now delivered them using Varian’s Aria solution and storing imaging within Aria.

Due to the time, costs and difficulties associated with retrieving legacy RT data stored within proprietary file formats, decisions were made as to the inclusion of radiotherapy data based on the **relevance of the treatment to modern and current treatment techniques, the ease of data retrieval and the paucity of the data.**

For this reason, only patients treated between 2011-2017 using IMRT and VMAT had their full datasets retrieved from legacy radiotherapy planning system, Monaco. No patient datasets were retrieved from XiO beyond those which had already been retrieved for clinically relevant situations for individual patient care.

This led to the exclusion of all palliative RT datasets before 2017 as well as cases treated with 3D-CRT. This primarily excluded the RT datasets of T1-T2 larynx patients treated radically pre-2017.

The datasets available for inclusion within the database are shown in Figure 2. Of the 2895 patients, 314 never received Head and Neck Cancer RT and 166 patients had their first course of radiotherapy before the introduction of inversely planned IMRT to the Trust which was defined as the cohort start date for radiotherapy data retrieval (March 2011). The majority of excluded patients having radical treatment were larynx patients (n=84/104).

***Figure 2: Patients eligible for radiotherapy data retrieval into RTDLHN.***

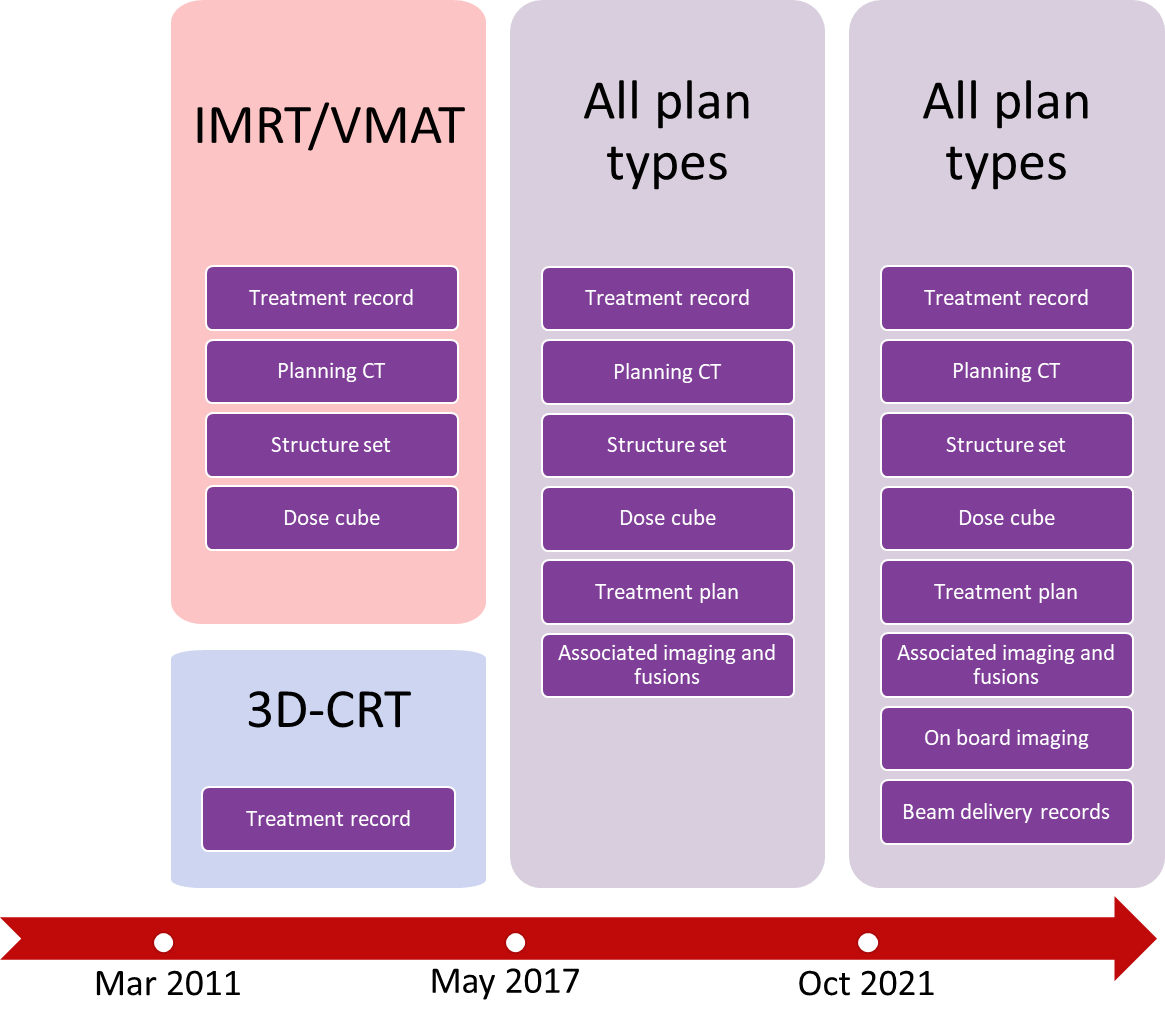
### Data selection

In addition to the above patient selection criteria, data selection criteria was also employed based on the **relevance of the data to modern and current treatment techniques, the ease of data retrieval and the paucity of the data.**

This excluded on-board imaging from before October 2021 when we moved to a new record and verify system, Aria, due to the complexities involved in data retrieval.

The data retrieved for each patient epoch is summarised in Figure 3.

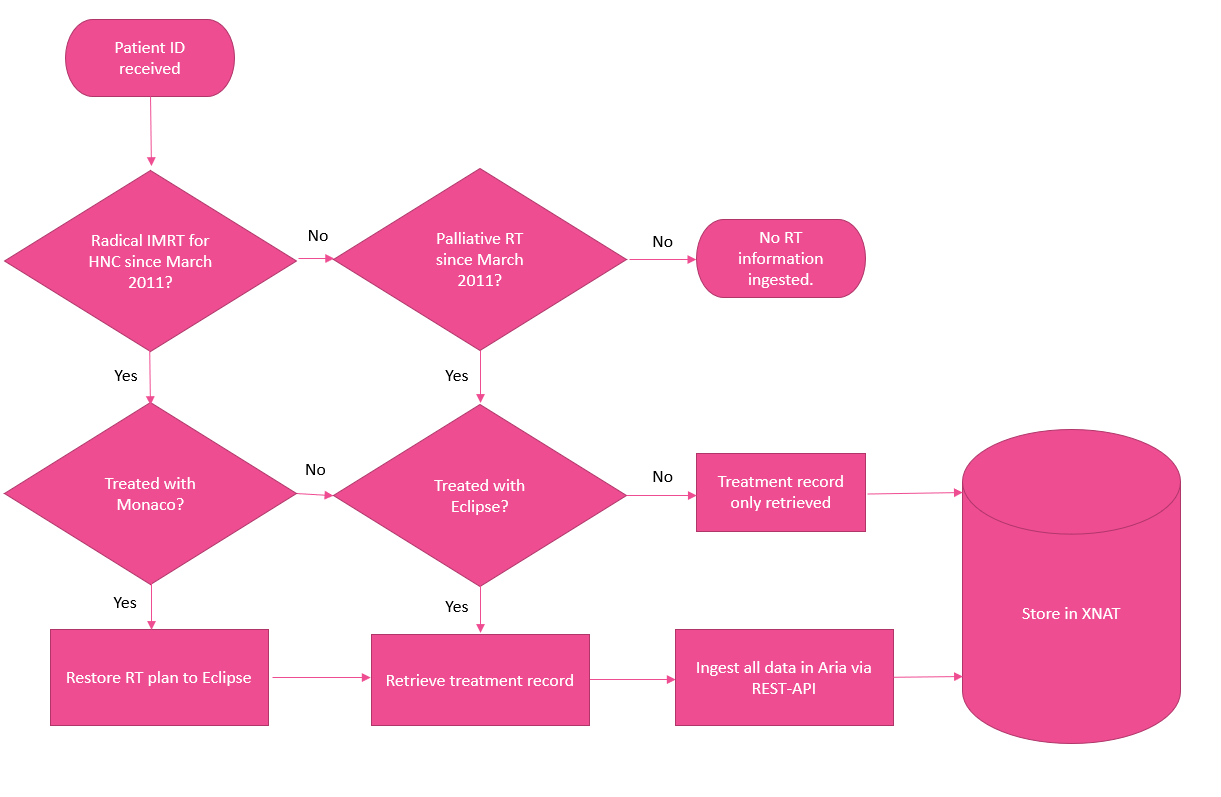
Of note, the treatment plan (by which we define the RTPlan file which contains the technical beam delivery control points such as leaf/jaw positions) is not available for patients pre-Eclipse, this is discussed in detail below.



***Figure 3: Data retrieved from each radiotherapy epoch. Between 2011-2017, 3D-CRT plans were created in XiO and IMRT/VMAT plans were produced within Monaco.***

### Data transfer process

The relevant legacy treatment plans from Monaco were first imported into Eclipse to facilitate ease of treatment planning should the patients return for additional radiotherapy, and to enable a one-step transfer process of all RT data into XNAT. A data flow diagram for this is shown in Figure 4. Treatment records were taken from either Mosaiq or Aria depending on the epoch of radiotherapy.



**Figure 4: Radiotherapy data flow from Radiotherapy systems to XNAT.**

The number of patients having their 1st treatment course in each treatment planning system is shown in Figure 5 below. Patients who had treatment in both Monaco and Eclipse are considered Monaco patients here to account for the significant burden of additional work represented by their data transfer.

***Figure 5: Originating treatment planning systems for data transfer to XNAT.***

#### Monaco patients restored to Eclipse

Patients were systematically archived from the Monaco Treatment Planning System to maintain working memory availability of the program when in clinical use. The process for data retrieval is detailed in work instructions for retrieving patients from Monaco but the basic steps are outlined here for appreciation of the amount of work involved. The processes are slightly different between IMRT and VMAT and so they are detailed separately.

During the period of retrieving datasets from Monaco into Eclipse, MRNs replaced patient IDs. During retrieval, each patient was checked on EPIC to see if they had an MRN. Patients with MRNs were relabelled with their MRN for use in Eclipse in case they come back for future treatment, to ensure patient records are not duplicated for safety reasons.

Of the 932 patients having their first treatment course in Monaco, there were n=578 IMRT plans and n=354 VMAT plans.

For patients replanned during the course of treatment, care was taken to include all the radiotherapy planning data of both the original plan and any replans. This was performed by checking the monitor units (MU) delivered each treatment fraction compared to the treatment plan. The number of fractions a patient received on each plan was carefully annotated onto the radiotherapy record description associated with each patient to ensure the accuracy of any future dose data mining.

|  |  |  |
| --- | --- | --- |
| Process | IMRT | VMAT |
| Locate archive location in pdf using patient ID | Checkmark | Checkmark |
| Locate archive location in folder | Checkmark | Checkmark |
| Copy patient folder to Monaco working directory | Checkmark | Checkmark |
| Open patient in Monaco | Checkmark | Checkmark |
| Export patient CT, structure set, treatment plan and dose cube to drive. | Checkmark | Checkmark |
| Import from drive to Eclipse | Checkmark | Checkmark |
| Check expected elements present and correct (CT, SS, dose cube) | Checkmark | Checkmark |
| Create dummy treatment field |  | Checkmark |
| Reimport dose cube |  | Checkmark |
| Check expected elements present and correct (CT, SS, dose cube) |  | Checkmark |
| Check imported plan against Mosaiq delivered fields | Checkmark | Checkmark |
| Check patient ID for new EPIC MRN (for work post October 2023) | Checkmark | Checkmark |

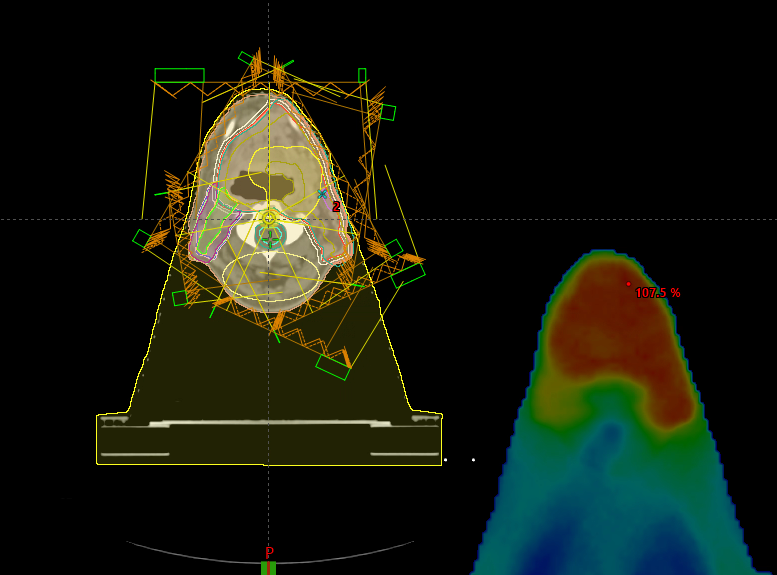
##### Transfer of Monaco patients treated with IMRT (March 2011- ~2014)

Patients treated with IMRT (~7 static beams with modulated fluence) were exported from Monaco to disk and then imported into Eclipse. They were visually inspected to ensure the CT, structure set and dose cube appeared as expected and the delivered MU in Mosaiq was compared to those of the plan to ensure the correct plans had been located.

###### Known IMRT data quality issues

Dose cube shift

Some patients (n=25) from the very beginning of the use of IMRT had a misalignment, when transferred to Eclipse, between their dose cube and their CT and structure set. An example is shown in Figure 6.



***Figure 6: Dose cube misalignment error post-transfer to Eclipse.***

This misalignment results in incorrectly calculated dose statistics. Work was undertaken to realign these patients both manually and in an automated fashion using manipulation of frame of reference DICOM tags which was only successful on n=1 patient; however, n= 24 patients resisted realignment. These patients have been marked as having a known error but should a global realignment method be found; their data would be easily available. This is particularly relevant for projects with only small numbers of patients to draw from. It also demonstrates the value of visual inspection of data for machine learning purposes.

Miscellaneous

In addition to the dose cube shift, n=5 patients treated with IMRT suffered from known data quality issues, n=2 patients had missing replans (but have their original data available), n=1 patient cannot be located entirely in any archiving system, n=1 is unable to be exported from Monaco due to a software bug and for n=1 patient there was a known error during treatment and so reconstructive work to determine the dose the patient received would be required which necessitates more Monaco technical expertise than the team possessed.

|  |  |
| --- | --- |
| **n IMRT plans not uploaded to XNAT** | **Reason** |
| 24 | Dose cube misalignment |
| 2 | Replans missing but original plan available |
| 1 | All plans and replans cannot be located entirely in any archiving system |
| 1 | Software bug in Monaco prevents export of patient |
| 1 | Known error during treatment requiring reconstruction of delivered dose |

Table 1: (n=34) IMRT plans not transferred to XNAT

#### Transfer of Monaco patients treated with VMAT (2014-2017)

Monaco VMAT plans could not be ingested to XNAT via Eclipse due to the coding of VMAT beam arrangements within the RTPlan.. Eclipse cannot deal with an beam intended to travel both clockwise and counter-clockwise as it expects 2 separate beams to perform this motion. For this reason, a dummy treatment field had to be created in each patient record to apply the dose cube on. This involved the creation of a new course, a new plan, selecting a static 6X treatment field and creating a reference point with its limits. The dose cube could then reimported and attached to this newly created field. This ensured the dose cube was attached to the patient and could be isplayed and manipulated within Eclipse.

For patients with multiple VMAT treatment plans, separate dummy plans needed to be created for each structure set/CT and the dose cube applied.

###### Known VMAT data quality issues

Lost Monaco B patients

There was a large corruption of the Monaco database in 2016 leading to the majority of patients from 2015 and 2016 not been archived according to the above procedure. The patient data was likely to have been archived on XVI however, and so the plans could be recalculated using a research instance of Monaco. This research instance of Monaco was, however, not commissioned for clinical use, and run version 5.1 of Monaco instead of version 5.0. The beam model is the same for both versions, hence it could be used to perform equivalent, non-clinical calculations. This was a lengthy procedure taking upwards of 1 hour per patient.

Firstly, the XVI archives were interrogated to see if the lost patients existed within them. After finding the location of the archive, the files were copied into Conquest, an open source DICOM server allowing export of files producing a .dir file which enables Monaco to “see the files”, and then sent to Monaco. Within Monaco, the patient could then be imported. With the patient imported containing a CT, structure set and plan file, a couch structure needed to be added and the densities assigned; then the calculation parameters set with a grid spacing of 0.25cm, dose deposition to water and statistical uncertainty per calculation set to 0.7%. Calculation then took up to 25 minutes with no progress indicated. After calculation, the patient was saved, exported to file and then imported into Eclipse. The majority of these patients were VMAT patients and so a dummy plan then had to be created to enable the import of the dose cube into Eclipse.

Some patients did not have a body contour present and so recalculation was not possible. It was estimated that to draw the body contour would take an additional hour per plan per patient and it was therefore decided not to retrieve these cases.

Of 128 patients identified as Monaco B patients, 76 were fully retrieved, 6 were partially retrieved and 3 were ineligible for inclusion within the radiotherapy dataset initially. Unsuccessful retrievals (n=44) have been indicated in the radiotherapy dataset to illustrate their possible inclusion in future datasets if they are of particular interest.

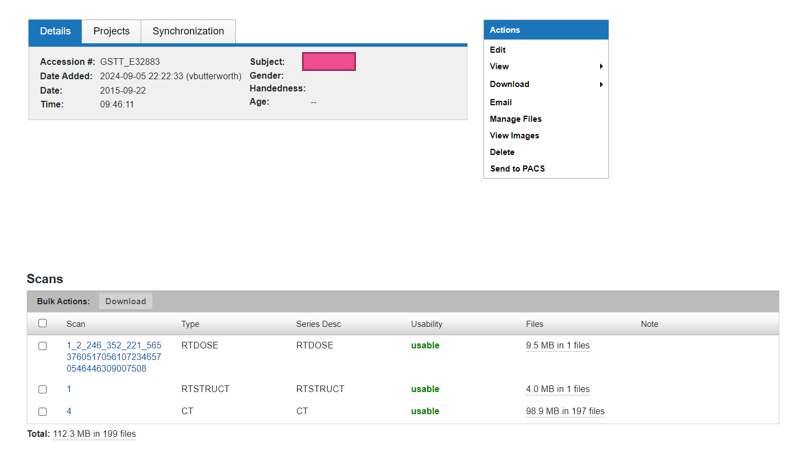
***Figure 7: Monaco B patients retrieval summary***

Unable to be located

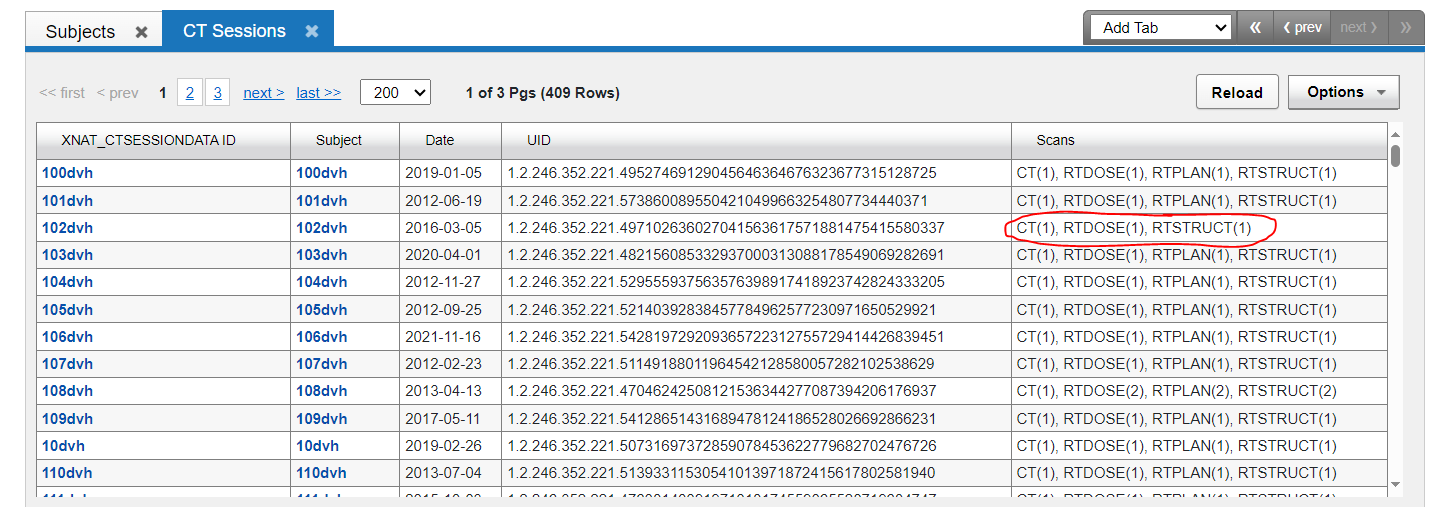
As detailed above, there is no guarantee that all Monaco B patients are within the XVI archive. Additionally, n= 1 full patient and n=1 replan from the end of the clinical service (March 2017 onwards) appear to not have been archived from Monaco and have been lost from the working directory and are therefore not available.

Incorrectly reconstructed

During the course of the transfer of radiotherapy data to XNAT, a subset of reconstructed VMAT plans failed in transferring the dose cube. This is because they had been imported into Eclipse without a dummy treatment field to attach the dose cube to which meant that XNAT could not “see” the dose cube. These patients were manually transferred to XNAT from Eclipse rather than using the below detailed scripted pull. For future projects using these patients (n=~167), researchers will have to create a dummy treatment field to view the dose cube in Eclipse. This will not affect research taking place outside of the TPS because all dose cubes are intact.



In XNAT these patients appear with only 3 DICOM scans attached to their radiotherapy dataset, CT, RTSTRUCT and RTDOSE and so it is obvious that plan reconstruction needs to occur. This can be observed also in the scan summary of all sessions available per patient.



#### Total data retrieval

Considering all of the above, the availability of radiotherapy datasets is summarised below. During the course of this work, no errors in Eclipse datasets were identified and so these are assumed to be intact and to date no Eclipse data corruption is known to have occurred. Of n=2170 patients available for inclusion, n=2071 patients were successfully retrieved. Of the remaining n=99 patients with known data quality issues, these can be summarised as dose cube misalignment (n=24), having missing full datasets (n=52) and having partial missing datasets (n=23) where the patient was known to have been replanned but either the replan or the original plan could not be located.

***Figure 8: Known data quality issues compared to successful dataset retrievals.***

A graph of missing data against time is shown in Figure 9 below. Clear peaks at the introduction of IMRT due to the dose cube misalignment errors and for 2015-2016 patients due to the Monaco B database corruption.

***Figure 9: Missing radiotherapy datasets per year***

CBCTs are only available for patients from Aria which was introduced in October 2021, the number of available CBCTs for datasets is therefore much smaller than full radiotherapy datasets. This is shown in Figure 10.

### Data transfer into XNAT

With all transferable radiotherapy datasets in Eclipse, datasets were transferred to XNAT using custom XNAT scripts ([XNAT/csc\_xnat at main · GSTT-CSC/XNAT](https://github.com/GSTT-CSC/XNAT/tree/main/csc_xnat)). Patient IDs were provided to Clinical Scientific Computing (CSC). CSC provide a list of study descriptions, accession numbers, dates of what XNAT can see within Aria and session names are constructed based on the data types returned for naming in XNAT. Names of sessions were constructed following the below pattern:

|  |  |  |  |
| --- | --- | --- | --- |
| RT session type examples | Labelling convention | Example | Data element examples (“scans”) stored as dicom files |
| Pre-treatment imaging | PatientAriaID\_RT\_year\_AccessionNumber | 3456789\_RT\_2017\_RJxxxxxxxxxx | Diagnostic MRI or CT or PET-CT scans, regs. |
| RT session | PatientAriaID\_RT\_year | 3456789\_RT\_2017 | CT, RTSS, RTDose, RTPlan and for post Aria patients; CBCTs, CBCT SS and regs and RTRecords. |
| RT session replan | PatientAriaID\_RT\_year\_REPLAN/RESCAN | 3456789\_RT\_2017\_REPLAN | CT, RTSS, RTDose, RTPlan, reg, CBCTs, CBCT SS and regs and RTRecords. |
| RT session (multiple within 1 year e.g. palliative mets) | PatientAriaID\_RT\_year\_BodyPartTreated | 3456789\_RT\_2017\_SPINE  3456789\_RT\_2017\_WHOLEBRAIN | CT, RTSS, RTDose, RTPlan |
| On-treatment CBCT (pre-Aria) | PatientAriaID\_RT\_year\_CBCT | 3456789\_RT\_2017\_CBCT | CBCT scan, reg, RTSS |
| Patients from Monaco RT session | PatientAriaID\_RT\_year  OR  PatientAriaID\_RT\_H\_N\_Monaco OR PatientAriaID\_dateCT\_Tmt\_IMRT | 3456789\_RT\_2017  3456789\_RT\_H\_N\_Monaco  3456789\_18-07-2012\_Tmt\_IMRT | CT, RTSS, RTDose, dummy treatment field |

***Table 1: Naming conventions in XNAT and data elements transferred where CT refers to planning CT, RTSS= Radiotherapy Structure set, RTDose= RT Dosecube file, RTPlan= RT Plan file, reg = registrations to planning CT, CBCT SS= Structure set associated with CBCT, RTRecords= beam delivery records produced by Aria.***

The session names were returned to CSC in a .csv file where a custom script was used to ingest the indicated sessions with the given names. Imaging sessions were stored under the patient’s identifier within XNAT (NHS number).

All verification studies were ignored. This was a 2-fold decision based on verification plans being unlikely to be of value in clinical research and the simplicity of recreating them in addition to technical considerations.Verification plans all reference the same CT scan of the verification phantom and identical CTs are not permitted to be stored in multiple locations in XNAT leading to multiple corruptions and conflicts when we attempted to ingest verification plans.

XNAT ingests all available files attached to an imaging session in Aria. This means that all CBCTs for post-Aria patients are attached to the imaging session containing the original planning CT, structure set and dose cube (likely to be the things most of interest to a researcher). This means careful selection of data must take place when giving data to researchers. This is covered in the downloading data from XNAT SOP.

A .csv file detailing the radiotherapy information retrieved and expected in XNAT was produced, this is attached to the RT-HaND\_I data lake as a .csv file in XNAT, this also features the descriptions of how many fractions patients received on each plan in the event of replans and any known errors in data retrieval.

# Discussion

We have created a clinically annotated, carefully curated, data lake of medical imaging data of 2895 consenting H&N patients for the purposes of federated machine learning containing over 2071 quality assured patient radiotherapy datasets. Some patients (n= 175) will have had multiple courses of radiotherapy (as of October 2024) and so n= 2071 represents the minimum number of usable radiotherapy datasets retrieved.

This is a large dataset containing multi-dimensional radiotherapy data to answer many multi-faceted radiotherapy related questions. This SOP highlights the primary challenges involved in curating large volumes of retrospective and prospective radiotherapy (RT) data, including the necessary data cleaning, checking and structuring to ensure it is ready for ML applications. The process involved carefully selecting patient cohorts based on inclusion/exclusion criteria that prioritize relevance to current treatment approaches and the ability to retrieve data accurately. Overcoming these challenges required a standardised method to resolve technical obstacles related to legacy data formats, database corruption issues, and varying data storage protocols over time and intense data checking protocols to ensure the quality of the data. Model accuracy is only as good as the data quality and so feeding machine learning models data containing errors and noise will lead to unreliable results hence the high upfront cost associated with building such a database.

The SOP details the rigorous quality control measures applied to ensure data integrity, such as visually inspecting imported dose distributions to prevent inaccuracies in dose statistics, especially for early IMRT datasets. Despite efforts to transfer all necessary data accurately, known deficits remain, primarily due to missing data for certain treatment epochs and corrupted datasets from legacy systems. These issues underscore the complexity involved in co-locating historical data from disparate systems into a unified database. They emphasise the importance of developing more automated and standardized data retrieval methods, especially for older or legacy systems, like Tomotherapy. Furthermore, this experience highlights the critical need to consider the ease of data retrieval and system interoperability when procuring new technologies, particularly for those systems likely to reach end-of-life within a clinical workflow. Building such considerations into purchasing decisions upfront can reduce data management challenges and support continuity of care and research over time.

This is a large HNC radiotherapy database. Compared to RADCURE [15], one of the most extensive open accessible HNC imaging datasets containing 3346 patients, our radiotherapy dataset is deeper in scope containing dosecubes to enable prognostic modelling leveraging dosimetric factors. Additionally, we have multimodal radiotherapy imaging beyond just CT images, these additional imaging modalities may provide information beyond patients’ clinical indicators and their disease characteristics beyond the anatomical and density information provided by CT. RADCURE however have standardised the nomenclature for individual organ at risk and GTV contours which would be a valuable extension in scope for us in future.

# Conclusions

We have created a large radiotherapy repository of historic and current radiotherapy treatment plans to facilitate audit and research. Future work will focus on the automation of acquiring new patient treatment records from Aria via the AURA database server to minimize manual curation for prospective patients and courses of treatment.

# References

1. Cancer Research UK (2023) Cancer statistics for the UK

2. Dean JA, Wong KH, Welsh LC, et al (2016) Normal tissue complication probability (NTCP) modelling using spatial dose metrics and machine learning methods for severe acute oral mucositis resulting from head and neck radiotherapy. Radiotherapy and Oncology 120:21–27. https://doi.org/10.1016/j.radonc.2016.05.015

3. Dean J, Wong K, Gay H, et al (2018) Incorporating spatial dose metrics in machine learning-based normal tissue complication probability (NTCP) models of severe acute dysphagia resulting from head and neck radiotherapy. Clin Transl Radiat Oncol 8:27–39. https://doi.org/10.1016/j.ctro.2017.11.009

4. Humbert-Vidan L, Patel V, Andlauer R, et al (2022) Prediction of Mandibular ORN Incidence from 3D Radiation Dose Distribution Maps Using Deep Learning

5. Di Maio M, Perrone F, Conte P (2020) Real-World Evidence in Oncology: Opportunities and Limitations. Oncologist 25:e746–e752. https://doi.org/10.1634/theoncologist.2019-0647

6. Hwang E, Burnet NG, Crellin AM, et al (2022) A Novel Model and Infrastructure for Clinical Outcomes Data Collection and Their Systematic Evaluation for UK Patients Receiving Proton Beam Therapy. Clin Oncol 34:11–18. https://doi.org/10.1016/j.clon.2021.09.010

7. Langendijk JA, Orecchia R, Haustermans K, et al (2018) Prospective data registration and clinical trials for particle therapy in Europe. Radiotherapy and Oncology 128:9–13. https://doi.org/10.1016/j.radonc.2018.06.001

8. Boehm KM, Khosravi P, Vanguri R, et al (2022) Harnessing multimodal data integration to advance precision oncology. Nat Rev Cancer 22:114–126. https://doi.org/10.1038/s41568-021-00408-3

9. Eder J, Shekhovtsov VA (2021) Data quality for federated medical data lakes. International Journal of Web Information Systems 17:407–426. https://doi.org/10.1108/IJWIS-03-2021-0026

10. Batini C, Scannapieco M (2016) Data and Information Quality. Springer, pp 21–51

11. Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3:160018. https://doi.org/10.1038/sdata.2016.18

12. Marcus DS, Olsen TR, Ramaratnam M, Buckner RL (2007) The extensible neuroimaging archive toolkit. Neuroinformatics 5:11–33. https://doi.org/10.1385/NI:5:1:11

13. Beasley D, Eiben B, Doran S, et al (2019) Secure-DICOM-Uploader: A platform for anonymising and transferring imaging data from hospital sites to remote repositories

14. Kalendralis P, Traverso A, Shi Z, et al (2019) Multicenter CT phantoms public dataset for radiomics reproducibility tests. Med Phys 46:1512–1518. <https://doi.org/10.1002/mp.13385>

15. Welch ML, Kim S, Hope AJ, et al. RADCURE: An open-source head and neck cancer CT dataset for clinical radiation therapy insights. Med Phys. 2024;51(4):3101-3109. doi:10.1002/mp.16972