Relevant peer reviewed abstracts regarding XNAT construction

CRUK-ARR 2023

"Data-centric Artificial Intelligence and Cancer Research: Constructing a Real World Database using XNAT."

Victoria Butterworth, Dijana Vilic, Haleema Al Jazzaf, Isabel Palmer, Joshua Andriolo, Tania Avgoulea, Sarah Misson-Yates, Teresa Guerrero-Urbano.

BACKGROUND: The optimal performance of machine learning models and their generalisability is hinged on the quality of data used for model construction. The aim of the project was to set up a high quality scalable imaging and Radiotherapy (RT) Head and Neck (H&N) database with annotated clinical data. This multidisciplinary collaboration aimed to increase the speed and reduce the cost of research data curation by systematically acquiring patient data in line with data-centric artificial intelligence principles with special emphasis on patients' rights to privacy and data protection. It was intended that this database will facilitate Real World Evidence (RWE) reporting and enable evaluation of the relationship between radiomic and dosimetric parameters on tumour control probability and normal tissue complication probability that, linked to genomic data, will allow the design of clinical studies of personalised radiotherapy.

MATERIALS AND METHODS: XNAT is a powerful open source platform capable of storing and managing medical imaging and associated data. It provides import, archiving, processing 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. We have created a secure XNAT data lake of consenting H&N RT patients hosted by Clinical Scientific Computing team at GSTT. This data lake can be continuously updated with imaging data, Radiotherapy data and non-imaging data. A secondary XNAT has been established locally within Radiotherapy to host anonymised research projects, enabling investigators with access rights to download data from the main data lake for individual, ethics-approved studies.

RESULTS: DICOM node associations have been set up to the RT treatment planning system (Eclipse) and PACS (Sectra) to the data lake. A total of 3300 patients were identified for retrospective data retrieval, with an ongoing prospective data retrieval of 300 patients per year. We established protocols for the recovery and ingestion of legacy data, of which 800 patients have been identified and recovered. Data quality checking requirements have been defined and protocols for data anonymisation to facilitate research studies are in place. A set up system of general administration of data, users and user access has been established. Test datasets for functionality testing have been identified and over 8000 imaging sessions have been retrieved from clinical systems and stored in the XNAT H&N data lake so far, with non-imaging retrieval to commence soon.

CONCLUSIONS: We have created a secure and extensible imaging and H&N RT cancer database. This data lake of up to date data for H&N research will facilitate much quicker data access for future AI and research projects in H&N RT. The secondary XNAT provides safeguards for data access and control, securing the data and ensuring it is used in accordance with high standards of information governance. 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.

ESTRO 2024

2760

Data-centric AI and cancer research: constructing a research data access pipeline using XNAT

Victoria Butterworth[1], Dijana Vilic[1], Haleema Al Jazzaf[1], Thomas Young[2,3], Isabel Palmer[4], Tania Avgoulea[1], Josh Andriolo[1], Carole Creppy[1], Corla Routledge[2], Sarah Misson-Yates[1], Teresa Guerrero-Urbano[2,3]
[1]Guy's and St. Thomas' NHS Foundation Trust, Medical Physics, London, United Kingdom. [2]Guy's and St. Thomas' NHS Foundation Trust, Radiotherapy, London, United Kingdom. [3]King's College London, School of Cancer and Pharmaceutical Sciences, London, United Kingdom. [4]King's College London, School of Biomedical Engineering and Imaging Sciences, London, United Kingdom

Topic

Interdisciplinary: Other (topic of relevance for radiation oncology, NOT related to any other category)

Keywords

Real World Evidence

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 open-source project architecture 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.

Material/Methods

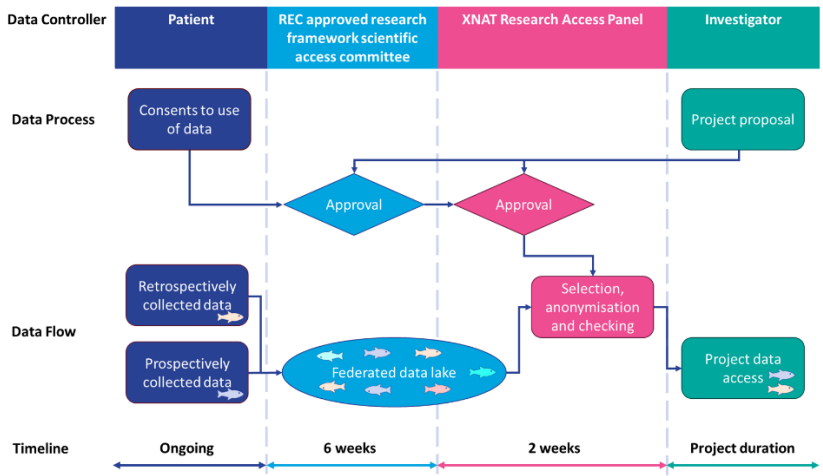
At our organisation, a valuable pre-existing Research Ethics Committee (REC) approved research framework (Reference: 18/NW/0297) is already in place for treated oncology patients which enables the use of clinical data for research [1]. All patients over the age of 18 years are eligible following their first visit for a diagnosis of active new or recurrent cancer and during consent to treatment they can opt-out of inclusion and their data being used for research purposes. A scientific access committee meets regularly to review applications to access the data with reference to scientific merit, study design and the applicant's resources.

XNAT is a powerful open-source platform capable of storing and managing medical images and associated clinical data. It provides import, archiving, processing, and secure distribution facilities. At

our institution, it forms part of the local secure enclave for the purposes of federated learning in AI projects. A secure XNAT data lake of consenting HNC patients' data that can be continuously updated with imaging, RT and non-imaging clinical data has been set up hosted by the organisation. This neat solution is independent of the electronic patient record provider and the radiotherapy vendor to enable incorporation of legacy radiotherapy and clinical data. A secondary XNAT (RT-XNAT) has been established locally within the Radiotherapy department to host only anonymised data for approved research projects. The XNAT Research Access Panel (XRAP) reviews approved ethics applications and provides access to an anonymised copy of the relevant requested data within RT-XNAT to the project investigators as well as access to the model production platforms hosted by the organisation's Clinical Scientific Computing group that leads on model development.

Results

2975 H&N patients treated since the introduction of Intensity Modulated RT to the clinic (2011) have been included within the federated data lake. In a first, data-mining phase, retrospective data from consenting patients was harvested from electronic patient records, picture archiving communications systems, RT planning systems and RT record systems for transfer to XNAT. The data lake can continuously be updated with additional imaging, RT and non-imaging clinical data and a prospective data-farming program is being set up underpinned by regular quality controls. Figure 1 shows the data flow diagram and access path for investigators wishing to use data from the federated data lake. A simplified system of governance is in place for request of the data and estimated timescales of response are included within Figure 1.



Conclusion

We have created a secure and extensible imaging and RT clinically annotated HNC database that will enable future AI research. Having a secondary XNAT to separately host research projects with anonymised patient datasets provides additional safeguards for data access and control, securing the data and ensuring it is used in accordance with high standards of information governance. This database and access protocol promises to be an exceedingly useful tool for research, revolutionising the time and cost associated with the production of machine learning models.

References

[1] 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.

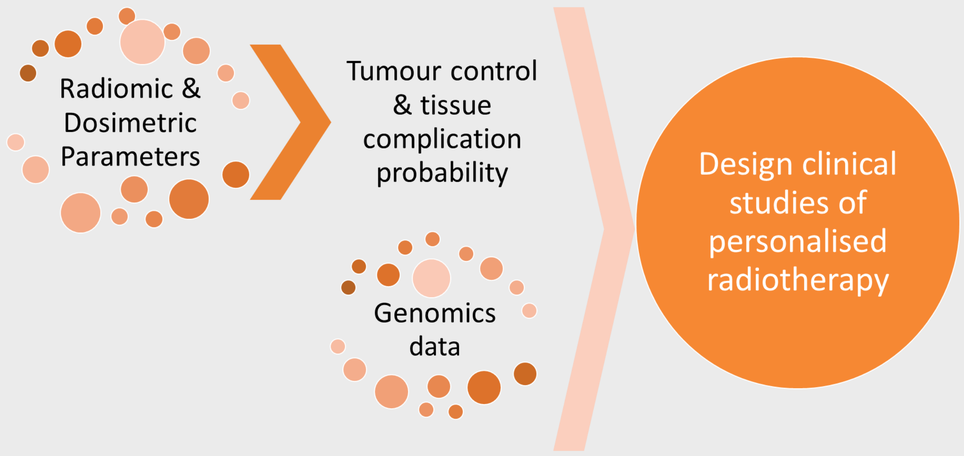
Data-centric Artificial Intelligence and Cancer Research: Constructing a Real World Database using XNAT

Victoria Butterworth, Dijana Vilic, Haleema Al Jazzaf, Isabel Palmer, Joshua Andriolo, Tania Avgoulea, Sarah Misson-Yates, Teresa Guerrero-Urbano

Introduction

The optimal performance of machine learning models and their generalisability is hinged on the quality of data used for model construction.

Systematically acquiring patient data in line with data-centric artificial intelligence principles will facilitate Real World Evidence (RWE) reporting and allow the design of clinical studies of personalised radiotherapy.



Aim

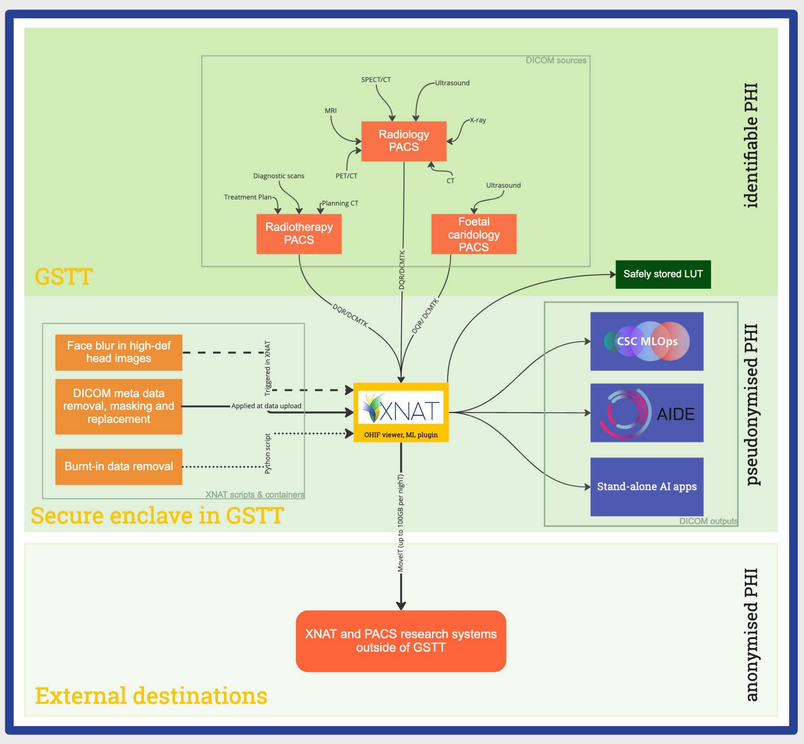
The aim of the project was to set up a high quality scalable imaging and Radiotherapy (RT) Head and Neck (H&N) database with annotated clinical data.

Database aims

- Increased speed of data curation
- Reduced cost of data curation
- Special emphasis on patients' rights to privacy and data protection

Method

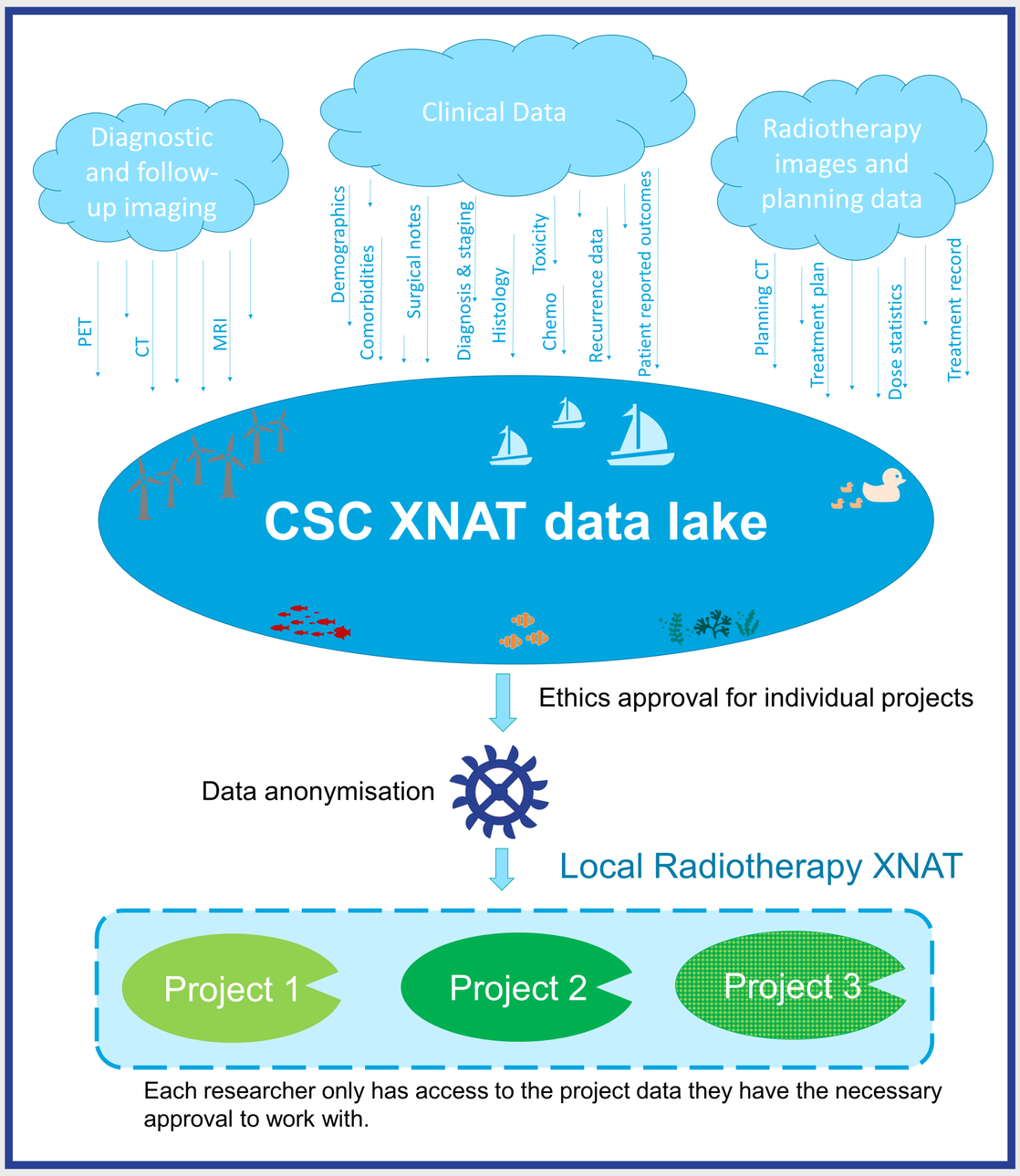
XNAT is a powerful open source platform capable of storing and managing medical imaging and associated data. It provides import, archiving, processing and secure distribution facilities. Within Guy's and St Thomas' NHS Foundation Trust (GSTT), it forms a part of the local secure enclave for the purposes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



We have created a secure XNAT data lake of consenting H&N RT patients hosted by Clinical Scientific Computing (CSC) team at GSTT. A secondary XNAT has been established locally within Radiotherapy to host anonymised research projects.

Results

DICOM node associations have been set up to the RT treatment planning system (Eclipse) and PACS (Sectra) to the data lake.



A total of 3300 patients were identified for retrospective data retrieval, with an ongoing prospective data retrieval of 300 patients per year..

Requirements defined:	Protocols established:	Work in progress:
Data anonymisation to facilitate research studies.	Recovery legacy data.	800 legacy patients identified and recovered.
Data checking requirements.	Ingestion legacy data.	Over 8000 imaging sessions retrieved and stored.
Test datasets for functionality.	General administration of data, users and user access.	Non-imaging data retrieval to commence soon.

Conclusion

We have created a secure and extensible imaging and H&N RT cancer database. This data lake of up to date data for H&N research will facilitate much quicker data access for future AI and research projects in H&N RT. The secondary XNAT provides safeguards for data access and control, securing the data and ensuring it is used in accordance with high standards of information governance. 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.

Our XNAT set-up enables machine learning model production to be:

- Safer
- Faster
- More efficient

Data-centric AI and cancer research: constructing a research data access pipeline using XNAT

Victoria Butterworth[1], Dijana Vilic[1], Haleema Al Jazzaf[1], Dr Tom Young[1,2], Isabel Palmer[2], Dr Tania Avgoulea[1], Josh Andriolo[1], Carole Creppy[1], Corla Routledge[1], Sarah Misson[1], Dr Teresa Guerrero Urbano[1,2]. Radiotherapy Department, Guy's and St Thomas' NHS Foundation Trust (GSTT) [1] and King's College London[2].

1. Introduction

- 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.

2. Aims

Database and data access pipeline aims

- Increased speed of data access
- Reduced cost of data curation
- Special emphasis on patients' rights to privacy and data protection

3. Methods

Patient has first visit for diagnosis of new or recurrent cancer

↓

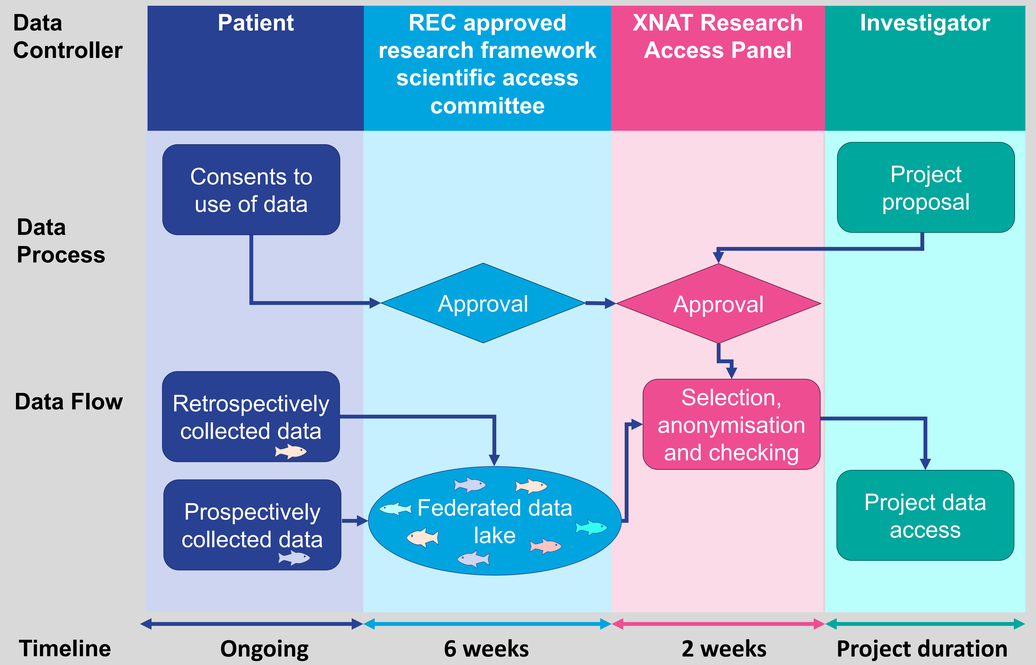
All patients over the age of 18 years eligible for inclusion

↓

Patients can opt-out during consent for treatment or any time thereafter

- At GSTT, a valuable pre-existing Research Ethics Committee (REC) approval enables the use of clinical oncology data for research [1].
- A **scientific access committee** meets regularly to review applications to access the data with reference to scientific merit, study design and the applicant's resources.
- **XNAT** is a powerful open-source platform capable of storing and managing medical images and associated clinical data. It provides import, archiving, processing, and secure distribution facilities.

4. Results



- A secure extensible XNAT federated data lake of **2895 consenting HNC patients**' imaging, RT and clinical data has been set up, hosted by the organisation.
- A secondary XNAT (RT-XNAT) has been established locally within the RT department to host anonymised data for approved research projects.
- **The XNAT Research Access Panel (XRAP)** reviews approved ethics applications and provides access to an anonymised copy of the relevant requested data within RT-XNAT to the project investigators.
- Retrospective data from consenting patients was harvested from electronic patient records, picture archiving communications systems, RT planning systems and RT record systems for transfer to XNAT.
- The data lake can continuously be updated with additional imaging, RT and non-imaging clinical data.
- A prospective data-farming program underpinned by **regular quality controls** is in progress.

5. Conclusions

- We have created a secure and extensible imaging and RT **clinically annotated HNC database** that will enable future AI research.
- A secondary XNAT hosting research projects with anonymised patient datasets provides **additional safeguards for data access and control**, securing the data and ensuring it is used in accordance with **high standards of information governance**.
- This database and access protocol promises to be an exceedingly useful tool for research, revolutionising the time and cost associated with the production of machine learning models.

Our XNAT set-up enables ML model production to be:

Safer | Faster | More efficient

References

[1] 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.