End-to-end framework for automated collection of large multicentre radiotherapy datasets demonstrated in a Danish Breast Cancer Group cohort

Refsgaard, Lasse; Skarsø, Emma Riis; Ravkilde, Thomas; Nissen, Henrik Dahl; Olsen, Mikael; Boye, Kristian; Laursen, Kasper Lind; Bekke, Susanne Nørring; Lorenzen, Ebbe Laugaard; Brink, Carsten; Thorsen, Lise Bech Jellesmark; Offersen, Birgitte Vrou; Korreman, Stine Sofia

Published in:
Physics and imaging in radiation oncology

DOI (link to publication from Publisher):
10.1016/j.phro.2023.100485

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Physics and Imaging in Radiation Oncology 27 (2023) 100485

Contents lists available at ScienceDirect

Physics and Imaging in Radiation Oncology

journal homepage: www.sciencedirect.com/journal/physics-and-imaging-in-radiation-oncology

End-to-end framework for automated collection of large multicentre radiotherapy datasets demonstrated in a Danish Breast Cancer Group cohort

Lasse Refsgaard [a,j], Emma Riis Skarsø [b,j], Thomas Ravkilde [c], Henrik Dahl Nissen [d], Mikael Olsen [e], Kristian Boye [f], Kasper Lind Laursen [g], Susanne Nørring Bekke [h], Ebbe Laugaard Lorenzen [i], Carsten Brink [i], Lise Bech Jellesmark Thorsen [a,c], Birgitte Vrou Offersen [a,b,c], Stine Sofia Korreman [b,c,j,*]

[a] Department of Experimental Clinical Oncology, Aarhus University Hospital, Denmark
[b] Danish Center for Particle Therapy, Aarhus University Hospital, Aarhus, Denmark
[c] Department of Oncology, Aarhus University Hospital, Aarhus, Denmark
[d] Department of Oncology, Vejle Hospital, University Hospital of Southern Denmark, Denmark
[e] Department of Oncology, Zealand University Hospital, Department of Clinical Oncology and Palliative Care, Næstved, Denmark
[f] Department of Oncology, Copenhagen University Hospital - Rigshospitalet, Copenhagen, Denmark
[g] Department of Medical Physics, Aalborg University Hospital, Aalborg, Denmark
[h] Department of Oncology, Copenhagen University Hospital – Herlev and Gentofte, Copenhagen, Denmark
[i] Laboratory of Radiation Physics, Department of Oncology, Odense University Hospital, Odense, Denmark
[j] Department of Clinical Medicine, Aarhus University, Aarhus, Denmark

ARTICLE INFO

Keywords:
Radiotherapy
DICOM
Data collection
Breast cancer
Automation
Big data
Data science

ABSTRACT

Large Digital Imaging and Communications in Medicine (DICOM) datasets are key to support research and the development of machine learning technology in radiotherapy (RT). However, the tools for multi-centre data collection, curation and standardisation are not readily available. Automated batch DICOM export solutions were demonstrated for a multicentre setup. A Python solution, Collaborative DICOM analysis for RT (CORDIAL-RT) was developed for curation, standardisation, and analysis of the collected data. The setup was demonstrated in the DBCG RT-Nation study, where 86% (n = 7748) of treatments in the inclusion period were collected and quality assured, supporting the applicability of the end-to-end framework.

1. Introduction

Big data and data science methods have the potential to accelerate the development of radiotherapy (RT) by acting as a supplement to the traditional translational research chain [1]. To take advantage of this potential, large-scale studies must move beyond binary registration of RT or prescribed dose and fractionations only and instead include the full exposure data (images, structure sets, treatment plans and 3D dose distributions) available in the Digital Imaging and Communications in Medicine (DICOM) format [2,3].

Large DICOM datasets also play a major role in the development of machine learning (ML) technology, which is rapidly finding its way into research and clinic. Unavailability of vendor provided functionality for bulk data exporting, needed to provide diverse training data [4], is however a hindrance.

In a recent survey 69% of respondents reported that they were either using or planning to use ML algorithms, naming the need for larger multicentre databases among the top priorities for going forward [5].

This need can be met by prospective data collection in clinical studies; however, it can be very time consuming for large datasets [6]. A multicentre collaborative effort to implement local methods for bulk DICOM data extraction would make this process faster and enable learning from archived treatment data, but may also increase the need for data curation, as data is not assessed on an individual level.

The variability and conformality of datasets depend on the extent of cross-centre collaboration and guideline implementation [7]. This is

* Corresponding author at: Dansk center for Partikel Terapi, Aarhus Universitetshospital, Palle Juul-Jensens Boulevard 25, indgang B3, 8200 Aarhus N, Denmark.
E-mail address: stine.korreman@oncology.au.dk (S.S. Korreman).

https://doi.org/10.1016/j.phro.2023.100485
Received 28 April 2023; Received in revised form 18 August 2023; Accepted 18 August 2023
Available online 25 August 2023

especially true for non-protocol treatments, which represent most of the available data. To address the task of curation, standardisation, and analysis of DICOM files, a vendor-agnostic tool is needed. Tools with standardisation capabilities exist, but these are either single purpose like nomenclature standardisation [8], focused on dose analysis such as the DVH Analytics package [9] or not open source like the DcmCollab system [10]. While not made for explorative data curation, a system like DcmCollab which focuses on storage, security and GDPR compliance, could however be used as a storage solution after the dataset has been curated. Though DICOM image data is traditionally stored in a Picture Archiving and Communications System (PACS), the widespread adoption of PACS in RT has been foiled due to several issues [11], making an RT-specific system a more suitable choice as the final step of an end-to-end framework.

In this technical note, we present and discuss the implementation of an end-to-end framework for providing large multicentre DICOM-RT datasets. This includes implementing multicentre bulk DICOM data extraction solutions and developing a solution to handle curation, standardisation and analysis of large DICOM-RT datasets prior to permanent storage. The setup is demonstrated in a case study (*Danish Breast Cancer Group (DBCG) RT Nation study)* and quality assurance (QA) is performed for dose-volume histogram (DVH)-parameter extraction.

2. Materials and methods

2.1. Defining a multicentre cohort

Patients can be identified using a database with treatment and patient characteristics and a central identification system such as a social security number or by each participating centre based on local registration of treatment and patient characteristics. To ease the subsequent data curation and analysis, it is advisable to predefine a set of inclusion criteria to limit and streamline the extent of the exported data.

2.2. Case study: DBCG RT-NATION data selection

In DBCG RT-Nation, patients were identified using the Danish Civil Registration System (CPR) identifiers [12] obtained from the DBCG database [13]. All patients who underwent surgery for early breast cancer in Denmark 2008–2016 with an indication for loco-regional RT according to DBCG guidelines were eligible. Treatment planning data was collected for the first breast cancer RT (and sequential boost if present). For adaptive RT, the treatment plan with the most delivered fractions was collected. If this information was missing, the first treatment plan was collected. Information on GDPR compliance can be found in the supplementary document.

2.3. Implementation of bulk DICOM data extraction

Treatment planning systems (TPS) rarely support bulk DICOM data

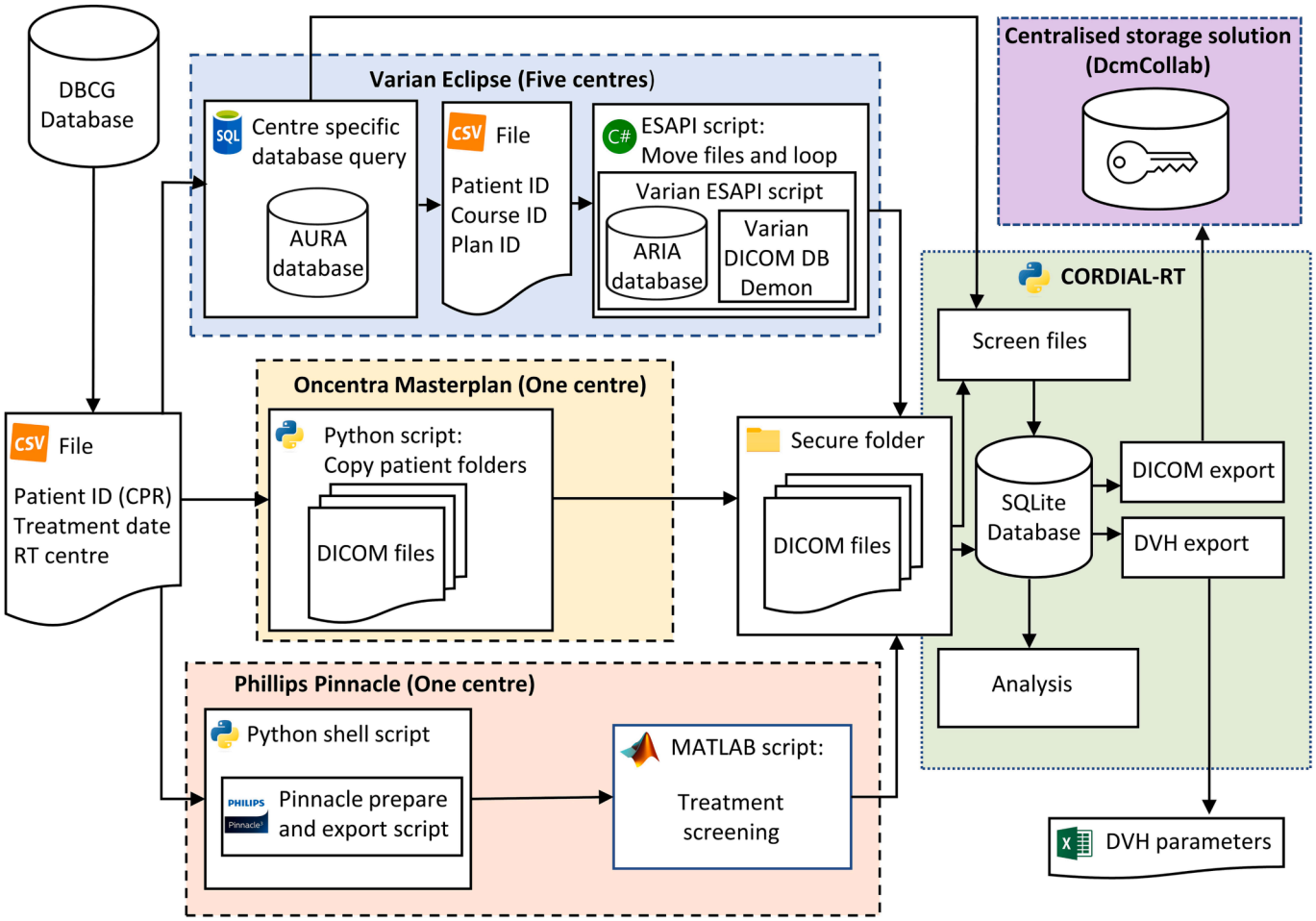

Fig. 1. Dataflow for the end-to-end framewok. DBCG: Danish Breast Cancer Group. CORDIAL-RT: Collaborative DICOM analysis for radiotherapy. CPR: Central Person Registry. AURA: Varian reporting solution. ARIA: Varian oncology information system. ESAPI: Eclipse Scripting API.

extraction as a standard solution, but do allow for scripting, which can be used to implement such extraction. Fig. 1 displays the complete end-to-end framework with solutions for the three TPS used in DBCG RT-Nation.

2.3.1. Eclipse (Varian Medical Systems)

In one of the Varian centres, a pilot project was carried out, implementing an application for automated batch export of DICOM files in the Eclipse Scripting Application Programming Interface (ESAPI), based on a script made available online by Varian [14]. We facilitated a workshop for all Varian centres where the application was shared for teaching purposes and later implemented in local variations by each Varian centre.

In the Varian centres, information on intended and actually delivered fractions of specific treatment plans was available in the ARIA database system. This information was used to find the dominant treatment plan (most fractions treated) and to filter out treatments that did not comply with inclusion criteria by using various automated methods, depending on local naming conventions and use of diagnosis codes.

2.3.2. Oncentra external beam (Nucletron B.V.)

The Oncentra centre had a DICOM file-based archiving system, which were organised in folders using the patient CPR, allowing easy extraction. No link between the archived plans and the number of fractions treated was implemented at the time, and the first plan was used if multiple plans were available, as all plans were planned with the full number of fractions by convention.

2.3.3. Pinnacle (Phillips)

The centre using Pinnacle implemented a local solution for an automated full DICOM data dump of their system. This solution was based on executing pinnacle scripts from a python shell script. A MATLAB script was used to select the relevant treatments.

2.3.4. Workload

Centres were surveyed to estimate the time spent on implementation and data extraction, which was compared between systems.

2.4. Collaborative DICOM analysis for radiotherapy (CORDIAL-RT)

After collecting and pseudonymising the DICOM files, a vendor-agnostic solution was needed to store and curate the large number of DICOM files. We developed the CORDIAL-RT solution, which consists of an SQLite database and a collection of functionalities made in python. CORDIAL-RT enabled scaling, summing and extraction of doses (based on Dicompyler-core python package [15]) as well as mapping of structure names and export of DVH data to a file and DICOM data to a centralised storage solution. A brief introduction to the solution is given in the document. The source code is available on Github [16].

2.5. Case study: DBCG RT-Nation data curation, standardisation and QA

CORDIAL-RT was used to curate and organise DICOM files into one treatment per patient. Treatments that did not fit the inclusion criteria were removed. If multiple dose files were associated with a treatment, the system would do automatic summing and save a new dose file representing the full treatment, provided the same image-set was referenced. In case of multiple image-sets, the dominant plan was used, and a scaling factor was added to the treatment and handled by the system. For sequential boost, doses were summed if the same image-set was used. All relevant structure names were categorised to a common name-set as defined in the DBCG Skagen trial 1 [17], based on the AAPM TG-263 report [18]. This was done by identifying the most frequent names using the Levenshtein distance to find similar named structures. The method was demonstrated on the ipsilateral lung, which was expected to be present in all treatments.

CORDIAL-RT was used to QA DVH-parameter extraction. Sample testing was done on a diverse subset of treatments (n = 20), comparing 87 dose and volume parameters for various structures, using the MATLAB CERR [19] package for independent validation.

Population dose QA was performed as a sanity check, by extracting the maximum treatment doses for all treatments and comparing the results to expected ranges. For a subset of treatments, the median target dose was also assessed.

3. Results

3.1. Implementation of bulk DICOM data extraction

Bulk DICOM-RT data extraction solutions were implemented for all centres. An estimated 6–7 workdays were spent developing the pilot solution for the Varian setup. For the other four Varian centres, estimated time spent was: 1–3 days implementing the automatic solution, 2–3 days developing and executing code for selecting and curating treatment data, and 1–2 min per patient for executing the automated export. For the Oncentra centre, a few hours were spent implementing the solution, three days for data curation and less than a minute for copying data for each patient. The Pinnacle centre could not estimate the time spent on this specific project, as it was done as part of a larger effort.

3.2. Case study: DBCG RT-Nation end-to-end demonstration

From the DBCG database, 9100 patients were identified. In total, DICOM data (~1.2 million DICOM files) for 8028 treatments (91%) was collected. In the screening processes before the data export, 246 treatments did not match the inclusion criteria and 826 eligible treatments could not be collected (Fig. 2). About half of the uncollected eligible treatments (n = 453), were from 2008 and non-retrievable due to loss of access to data storage. During the curation and standardisation process, 219 additional treatments were found to not fit the inclusion criteria. Furthermore, 334 treatments were either incomplete or inconclusive and could not be processed. In total, 7448 treatments (86%) were processed and included in the dose QA. From 2009 this was 90%.

In the sample test, all volume differences between CORDIAL-RT and MATLAB Cerr were <1% or <1 cm^3. All dose differences were <2% or <0.5 Gy (supplementary Table 1). The population dose QA showed that 9% of treatments had a maximum relative dose above 120% (supplementary Fig. 1). Of these treatments, 5% had a median target dose above 110% and 0.8% had a median target dose above 120% (supplementary Fig. 2). No treatments had a maximum relative dose below 100%. We identified 158 different names associated with the ipsilateral lung, one of which was present in all but one treatment. As a proof of concept, 200 treatments were successfully exported to the DcmCollab system.

4. Discussion

We demonstrated the feasibility of a national end-to-end framework for collecting large DICOM-RT datasets, exemplified in a curated national dataset of 7448 node-positive breast cancer patients treated 2008–2016. In 2009–2016, 90% of all loco-regional breast cancer RT treatments in Denmark were successfully collected and processed.

The 10% missing and inconclusive data from 2009 and later was caused by several factors e.g., centres not being able to identify all treatments automatically, some treatments not being exported with all the needed files and CORDIAL-RT not being able to process treatments with different structure-sets for primary and sequential boost plans. Despite the missing data, we were able to collect a dataset that is among the largest in radiotherapy containing full DICOM data. In comparison, the recently published CANTO-RT study from France included 3976 breast cancer patients [6].

The semi-automatic method used in CORDIAL-RT for categorising

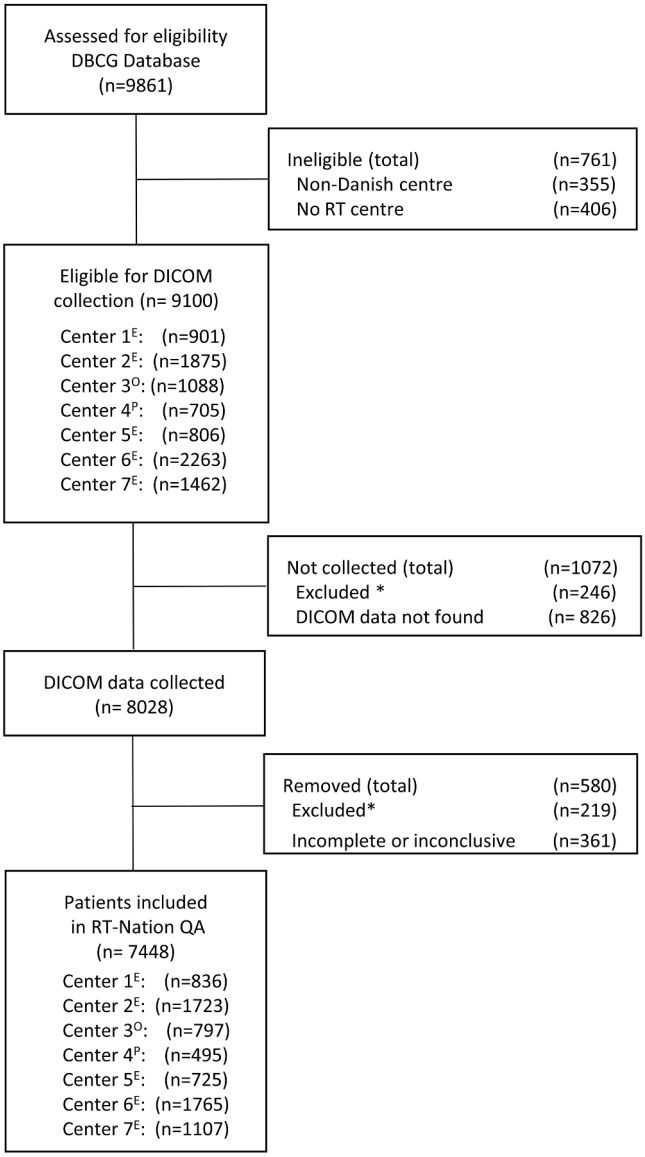

Fig. 2. CONSORT diagram. DBCG: Danish Breast Cancer Group. RT: Radiotherapy. QA: Quality Assurance. [E]Eclipse, Oncentra, [P]Pinnacle. *Inclusion criteria: Unilateral curative loco-regional breast cancer RT.

structure names required a fair amount of domain knowledge and time. A recent study [8] demonstrated an ML based method for standardising structure nomenclature on 1613 breast cancer patients with promising results. This could make the process faster and less subjective, however the method was only demonstrated in a single centre and was language dependant. With the rise of auto-segmentation tools, structure categorisation for dose evaluation will be less relevant in the future as re-segmentation of large cohorts will be more feasible. However, for evaluation of adherence to delineation guidelines, and for providing data to train auto-segmentation and ML models, clinical structure categorisation will remain important. The results of the independent dose validation test were comparable to what was found by a similar system [9] and the population dose QA did not point towards any problems.

Large data sets need to be easily and permanently available for optimal use. CORDIAL-RT is not intended as a permanent storage solution, but can supply curated and quality assured datasets to any system utilising standard DICOM communication. A closer integration of the methods presented here and the DcmCollab can be a viable way to deal with the challenges of maintaining large national datasets while adhering to increasingly strict data privacy regulations.

Sources of support

DCCC Radiotherapy – The Danish National Research Center for Radiotherapy, Danish Cancer Society (grant no. R191-A11526).

Danish Cancer Society (grant no. R204-A12389).

CRediT authorship contribution statement

Lasse Refsgaard: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization, Project administration. **Emma Riis Skarsø:** Methodology, Software, Validation, Data curation, Writing – review & editing. **Thomas Ravkilde:** Resources, Software, Writing – review & editing. **Henrik Dahl Nissen:** Resources, Validation, Writing – review & editing. **Mikael Olsen:** Resources, Data curation, Software. **Kristian Boye:** Resources, Data curation, Software, Writing – review & editing. **Kasper Lind Laursen:** Resources, Software, Writing – review & editing. **Susanne Nørring Bekke:** Resources, Data curation, Software, Writing – review & editing. **Ebbe Laugaard Lorenzen:** Resources, Data curation, Writing – review & editing. **Carsten Brink:** Resources, Data curation, Software, Writing – review & editing. **Lise Bech Jellesmark Thorsen:** Conceptualization, Writing – review & editing, Supervision. **Birgitte Vrou Offersen:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Stine Sofia Korreman:** Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.phro.2023.100485.

References

[1] Vogelius IR, Petersen J, Bentzen SM. Harnessing data science to advance radiation oncology. Mol Oncol 2020;14:1514–28. https://doi.org/10.1002/1878-0261.12685.

[2] Brink C, Lorenzen EL, Krogh SL, Westberg J, Berg M, Jensen I, et al. DBCG hypo trial validation of radiotherapy parameters from a national data bank versus manual reporting. Acta Oncol 2018;57:107–12. https://doi.org/10.1080/0284186X.2017.1406140.

[3] Thomsen MS, Berg M, Zimmermann S, Lutz CM, Makocki S, Jensen I, et al. Dose constraints for whole breast radiation therapy based on the quality assessment of treatment plans in the randomised Danish breast cancer group (DBCG) HYPO trial. Clin Transl Radiat Oncol 2021;28:118–23. https://doi.org/10.1016/j.ctro.2021.03.009.

[4] Field M, Hardcastle N, Jameson M, Aherne N, Holloway L. Machine learning applications in radiation oncology. Phys Imaging Radiat Oncol 2021;19:13–24. https://doi.org/10.1016/j.phro.2021.05.007.

[5] Brouwer CL, Dinkla AM, Vandewinckele L, Crijns W, Claessens M, Verellen D, et al. Machine learning applications in radiation oncology: current use and needs to support clinical implementation. Phys Imaging Radiat Oncol 2020;16:144–8. https://doi.org/10.1016/j.phro.2020.11.002.

[6] Sarrade T, Allodji R, Ghannam Y, Auzac G, Everhard S, Kirova Y, et al. CANTO-RT: one of the largest prospective multicenter cohort of early breast cancer patients treated with radiotherapy including full DICOM RT data. Cancers 2023;15:751. https://doi.org/10.3390/cancers15030751.

[7] Jensen M-B, Laenkholm A-V, Offersen BV, Christiansen P, Kroman N, Mouridsen HT, et al. The clinical database and implementation of treatment guidelines by the Danish Breast Cancer Cooperative Group in 2007–2016. Acta Oncol 2018;57:13–8. https://doi.org/10.1080/0284186X.2017.1404638.

[8] Haidar A, Field M, Batumalai V, Cloak K, Al Mouiee D, Chlap P, et al. Standardising breast radiotherapy structure naming conventions: A machine learning approach. Cancers 2023;15:564. https://doi.org/10.3390/cancers15030564.

[9] Cutright D, Gopalakrishnan M, Roy A, Panchal A, Mittal BB. DVH Analytics: A DVH database for clinicians and researchers. J Appl Clin Med Phys 2018;19:413–27. https://doi.org/10.1002/acm2.12401.
[10] Westberg J, Krogh S, Brink C, Vogelius IR. A DICOM based radiotherapy plan database for research collaboration and reporting. J Phys Conf Ser 2014;489:012100. https://doi.org/10.1088/1742-6596/489/1/012100.
[11] Shakeshaft J. PACS in radiotherapy. Clin Oncol 2010;22:681–7. https://doi.org/10.1016/j.clon.2010.06.010.
[12] Pedersen CB. The Danish Civil Registration System. Scand J Public Health 2011;39:22–5. https://doi.org/10.1177/1403494810387965.
[13] Christiansen P, Ejlertsen B, Jensen M-B, Mouridsen H. Danish Breast Cancer Cooperative Group. Clin Epidemiol 2016;8:445–9. https://doi.org/10.2147/CLEP.S99457.
[14] Keranen W. Scripting the Varian DB Daemon with ESAPI + DCMTK n.d. https://github.com/VarianAPIs/Varian-Code-Samples/wiki/Scripting-the-Varian-DICOM-DB-Daemon-with-ESAPI; 2019 [acessed 26 June 2023].
[15] Panchal A, Couture G, gertsikkema, Galler N, Hideki_Nakamoto, Hall DC, et al. dicompyler/dicompyler-core v0.5.5 2019. https://doi.org/10.5281/zenodo.3236628.
[16] Refsgaard L, Skarsø E. cordial-rt v 0.1 2023. 0.1 2023. https://github.com/Aarhus-RadOnc-AI/cordial-rt.
[17] Francolini G, Thomsen MS, Yates ES, Kirkove C, Jensen I, Blix ES, et al. Quality assessment of delineation and dose planning of early breast cancer patients included in the randomized Skagen Trial 1. Radiother Oncol 2017;123:282–7. https://doi.org/10.1016/j.radonc.2017.03.011.
[18] Mayo CS, Moran JM, Bosch W, Xiao Y, McNutt T, Popple R, et al. American Association of Physicists in Medicine Task Group 263: standardizing nomenclatures in radiation oncology. Int J Radiat Oncol 2018;100:1057–66. https://doi.org/10.1016/j.ijrobp.2017.12.013.
[19] Deasy JO, Blanco AI, Clark VH. CERR: A computational environment for radiotherapy research. Med Phys 2003;30:979–85. https://doi.org/10.1118/1.1568978.