Radiotherapy and Oncology 182 (2023) 109524

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Original Article

Tools for large-scale data analytics of an international multi-center study in radiation oncology for cervical cancer

Stefan Ecker [a,*], Christian Kirisits [a], Maximilian Schmid [a], Astrid De Leeuw [b], Yvette Seppenwoolde [d], Johannes Knoth [a], Petra Trnkova [a], Gerd Heilemann [a], Alina Sturdza [a], Kathrin Kirchheiner [a], Sofia Spampinato [c], Monica Serban [e], Ina Jürgenliemk-Schulz [b], Supriya Chopra [f], Remi Nout [d], Kari Tanderup [c], Richard Pötter [a], Nicole Eder-Nesvacil [a]

[a] Medical University of Vienna, Department of Radiation Oncology, Vienna, Austria; [b] University Medical Centre Utrecht, Department of Radiation Oncology, Utrecht, the Netherlands; [c] Aarhus University Hospital, Department of Oncology, Aarhus, Denmark; [d] Erasmus MC Cancer Institute, University Medical Center Rotterdam, Department of Radiotherapy, Rotterdam, the Netherlands; [e] McGill University, Department of Medical Physics, Montreal, Canada; [f] Tata Memorial Hospital, Department of Radiation Oncology, Mumbai, India

ARTICLE INFO

Article history:
Received 18 November 2022
Received in revised form 1 February 2023
Accepted 2 February 2023
Available online 9 February 2023

Keywords:
Data analytics
IGABT
Clinical trial monitoring
Cervical cancer

ABSTRACT

Purpose: To develop and implement a software that enables centers, treating patients with state-of-the-art radiation oncology, to compare their patient, treatment, and outcome data to a reference cohort, and to assess the quality of their treatment approach.

Materials and Methods: A comprehensive data dashboard was designed, which al- lowed holistic assessment of institutional treatment approaches. The software was tested in the ongoing EMBRACE-II study for locally advanced cervical cancer. The tool created individualized dashboards and automatic analysis scripts, verified pro- tocol compliance and checked data for inconsistencies. Identified quality assurance (QA) events were analysed. A survey among users was conducted to assess usability.

Results: The survey indicated favourable feedback to the prototype and highlighted its value for internal monitoring. Overall, 2302 QA events were identified (0.4% of all collected data). 54% were due to missing or incomplete data, and 46% originated from other causes. At least one QA event was found in 519/1001 (52%) of patients. QA events related to primary study endpoints were found in 16% of patients. Sta- tistical methods demonstrated good performance in detecting anomalies, with precisions ranging from 71% to 100%. Most frequent QA event categories were Treatment Technique (27%), Patient Characteristics (22%), Dose Reporting (17%), Outcome 156 (15%), Outliers (12%), and RT Structures (8%).

Conclusion: A software tool was developed and tested within a clinical trial in radia- tion oncology. It enabled the quantitative and qualitative comparison of institutional patient and treatment parameters with a large multi-center reference cohort. We demonstrated the value of using statistical methods to automatically detect implau- sible data points and highlighted common pitfalls and uncertainties in radiotherapy for cervical cancer.

Radiation oncology is an ever-evolving field that promotes innovation and imple- mentation of novel treatment techniques. Due to its multi-disciplinary setting, it offers the opportunity to collect large volumes and variety of data. However, as the complexity of available information increases, so do the challenges of processing, analyzing and validating it [1–3].

Data dashboards are software tools that offer a comprehensive, and interactive way to monitor and analyze large amounts of data. Their usage in healthcare has seen a dramatic surge in popularity due to the COVID-19 pandemic, that demonstrated their value for effective data analysis and reporting [4].

The recent rise of artificial intelligence (AI) incentivizes clinics to implement rou- tine data collection, in order to build advanced prediction models [5,6,2]. This also presents an opportunity to implement additional, independent quality assurance (QA) checks that automatically analyse the stream of incoming data for implausi- bilities and errors [7–9]. In this work we investigated the use of these data-driven approaches, to analyze data of patients treated with radiation therapy for cervical cancer, which is one of the most common malignancies in women globally [10,11].

Combined radiochemotherapy including magnetic resonance (MR) Image-Guided Adaptive Brachytherapy (IGABT) is considered state of the art treatment, achieving high levels of local control in

* Corresponding author at: Medical University of Vienna, Department of Radiation Oncology, Vienna, Austria.
E-mail address: stefan.ecker@meduniwien.ac.at (S. Ecker).

https://doi.org/10.1016/j.radonc.2023.109524
0167-8140/

patients with locally advanced cervical cancer (LACC) [12,13]. Over the past decades, the adoption of brachytherapy (BT) guided by three dimensional volumetric imaging has gained significant interest, especially in Europe, North America and Asia [14]. However, in part due to historical reasons, there are various clinical approaches to gynaecological BT with respect to dose prescription, fractionation and implant technique [15,16]. The introduction of a comprehensive target concept outlined by the GEC-ESTRO recommendations [17–20] and ICRU report 89 [15] enabled a common language for dose prescription and reporting.

Results of the EMBRACE-I study delivered compelling evidence for the efficacy of this approach, and support the clinical use of evidence-based dose objectives and pre- scription protocols for MR-IGABT [21]. These concepts, including advanced treat- ment techniques for external beam radiotherapy EBRT and BT are currently being investigated in the ongoing EMBRACE-II trial (NCT03617133) [14]. BT is regarded as a critical element for LACC treatment and a growing shift towards IGABT prac- tice has been observed [22,23]. However, IGABT is also a complex treatment, and a decrease in access to competent BT has been reported in parts of high-income countries as well [22].

Consequently, disseminating reproducible knowledge on optimal treatment is of paramount importance to the community, and advanced educational and training initiatives are needed to translate the improved outcome into clinical practice globally [23]. How- ever, centers practicing or adopting IGABT do not have a comprehensive way to compare their practice to experiences from other institutions. Based on the ongoing EMBRACE-II study, a software tool has been developed to enable centers treating LACC patients with radiotherapy, to compare their patient, treatment and outcome data to a reference cohort and assess the quality of their treatment approach. Ad- ditionally, as part of this tool, we investigated the utility of automatic anomaly detection methods, to monitor and ensure data quality. The automatic QA checks analyzed data for implausibilities, and flagged any unusual data points for review, which provided an additional level of quality assurance to support centers treating patients with IGABT.

Materials and methods

Patient cohort and data collection

Patients in EMBRACE-II were treated for LACC (FIGO stage IB-IVA, and nodal status according to TNM as N0 and N1). [14] Patients were treated according to the EMBRACE-II protocol. The treatment consisted of concomitant chemoradiation of 45 Gy external beam therapy with or without nodal boosts, followed by multifrac- tionated HDR or PDR brachytherapy with intracavitary and interstitial applicators. The total radiation dose from EBRT and BT was calculated as equieffective dose in 2 Gy per fraction (EQD2), using the linear-quadratic model with an $\frac{\alpha}{\beta}$ of 10 Gy for tumour, $\frac{\alpha}{\beta}$ of 3 Gy for OAR, and half time of repair ($T_{1/2}$) of 1.5 h [15]. The dose prescription protocol included a planning aim (soft constraint), for the total dose to the high risk CTV of D90 $\geq$ 90 Gy EQD2, and $\leq$ 65 Gy EQD2 for rectum and sigmoid, $\leq$ 80 Gy EQD2 Gy for bladder and $\leq$ 75 Gy for bowel. Dose limits (hard constraints) in case of failure to achieve the planning aims were applied in addition.

The study collected data from 51 centers. Recorded data included diagnostic in- formation based on clinical examination and imaging, dose-volume and treatment parameters for EBRT and MR-IGABT, and outcome data. Participating centers en- tered the information via an electronic case report form (eCRF) on a study website (https://www.embracestudy.dk). The data was stored in a centralized, anonymized database. At the time of writing, the study finished accrual with 1475 registered pa- tients. While the exact number of collected parameters depended on the individual case, approximately 600 data-points were collected per patient. 73 fields in the elec- tronic database included upfront logical tests and checks for valid numerical ranges. The developed software was only able to read, not write information to the study database. All participating centers in the study were given access to the software.

However, since the study was still ongoing during development, only 31 centers (1001 patients) were included in the analysis.

Software architecture

A custom software was designed using the statistical programming software R (ver- sion 4.0.2), with the Shiny package [24]. A web-server (CPU 2.6 GHz, 4 GB RAM) provided within the network of the study office's university was used to host the entire platform. 51 individualized dashboards, one for each center, were created. Views were limited to only the data from the particular institution compared to the entire study population. The overall structure of the project can be seen in Fig. 1.

Overall, the developed tool provided centers with an convenient way to access, ana- lyze, and explore their institutional data in comparison to the entire EMBRACE-II reference cohort, and get immediate feedback on data quality and protocol compli- ance. Users were able to access their individualized dashboards by logging in with a username and password via the web. Access to the tool was restricted to centers and researchers that participated in EMBRACE-II. Users accounts were created by the QA team and provided to participating centers. Based on the login information, the application returned the individualized view for the particular institution.

The overall tool structure was divided into five different sections: (i) *Overview* of data completeness, patient specific parameters and protocol compliance (ii) *EBRT* data (iii) *BT* data (iv) *Outcome* of disease and morbidity and (v) a list of *QA events*. As a large variety of parameters was collected, a pre-selection process was initially performed; In accordance with a multi-disciplinary expert team, a set of high-priority parameters to be included in the dashboard was defined. The goal was to offer users the most insightful information while at the same time keeping it interpretable and accessible. The final list of incorporated parameters can be found in Appendix Table A.1.

Custom functions were written to create graphs for 47 treatment and diagnostic parameters of interest. In general, continuous variables were summarized using Box- plots. Categorical variables were summarized using stacked bar-charts. Both offered direct comparison of the center-specific sub-cohort with the overall study population. In addition, statistical tests for significant differences between the two distributions were performed with each plot, using the Wilcoxon rank sum test for continuous vari- ables, and the Chi-squared test for categorical variables. A heatmap of all p-values obtained by these test can be found in Appendix Figure B.2.

The tool was designed to have a semi-automatic update process that ran on a weekly basis. The process involved extracting data from the EMBRACE-II study database and then cleaning and adding new variables through an extract, transform, load (ETL) process. During this process, scripts were run to analyze protocol compliance, calculate descriptive statistics, and assess data completeness for each patient.

Additionally, the ETL process included automatic data quality checks that scanned the database for inconsistencies and anomalies. Any findings were stored in a database and compared against the previous week's data. The tool's dashboard included an interactive table that allowed users to review the results of the identified

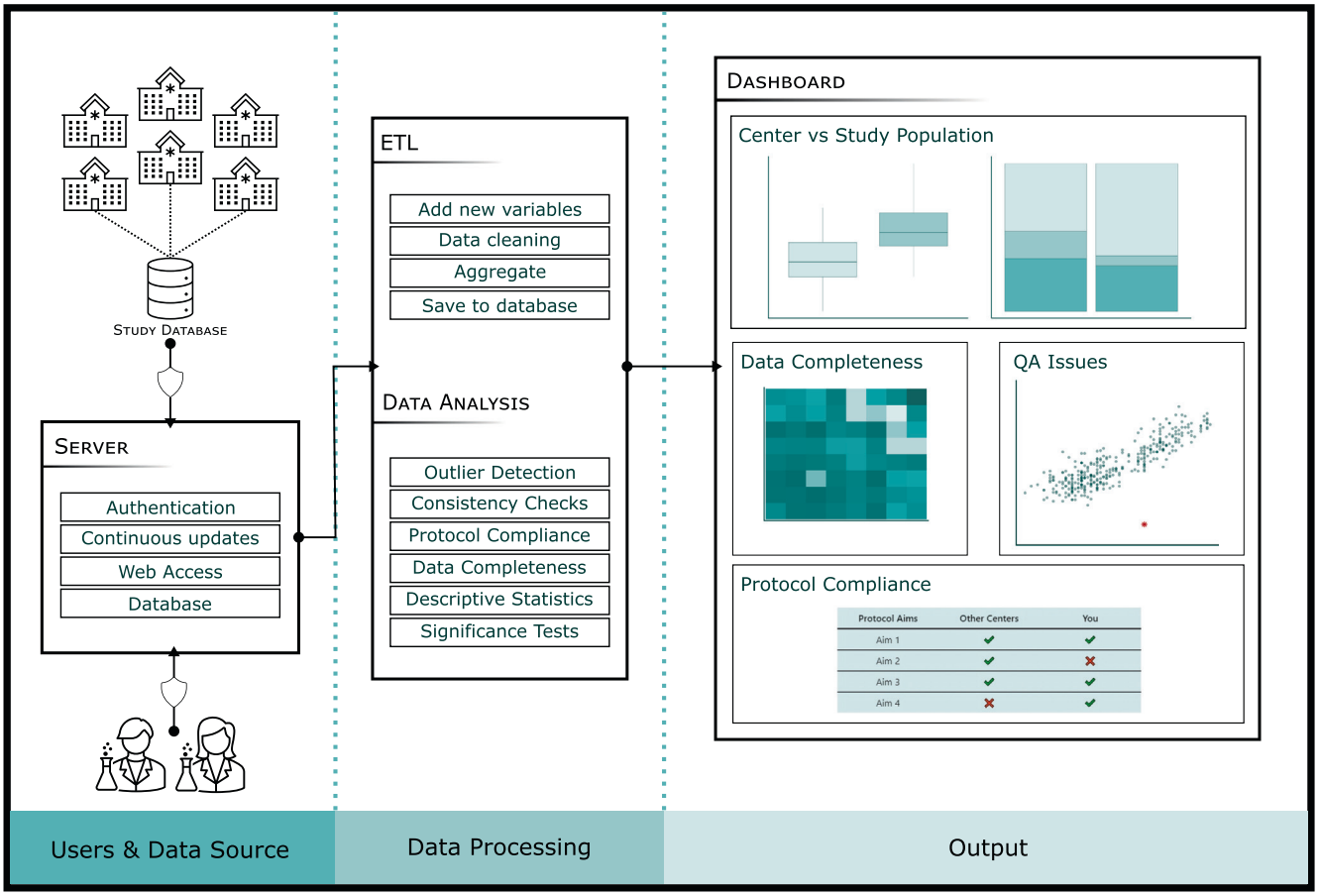


Fig. 1. Schematic overview of the project structure. Centers participating in the EMBRACE-II study enter data in a dedicated database. A software tool was developed on top of this data-source, that enables centers and researchers to compare their patient, treatment and out- come data to overall study population, and assess the quality of the treatment approach. A server hosts the platform, and enables secure access via the internet. The output after several data preprocessing and analysis steps, is a comprehensive data dashboard that contains additional automatic data QA checks. (ETL: Extract Transform Load, QA: Quality Assurance).

QA events. Each item in the table included a detailed description of the issue, the relevant patient ID, and the affected parameters in the study database. Users could also leave comments on each item for the study office. If any QA event was updated, both the initial and updated values were stored in the database and displayed in the table for reference. Detailed information about the algorithms follows in section 2.3.

Evaluation

To perform an evaluation of the tool's implementation and user experience, a short survey was conducted among participating centers. A link to an anonymous online survey was sent to the 31 EMBRACE-II centers that were included for analysis of this work. The survey closed after one month, returning responses from ten centers. The questionnaire consisted of eleven questions based on the System Usability Scale [25], which includes Likert-scale assessment of the most important features. The goal was to assess usefulness, usability and use-cases of the proposed software. In addition, participants were also asked to indicate how the system was used in their respective departments.

Data quality management

As part of the software platform, a database was set up to systematically collect find- ings ("QA events") from automatic data quality analyses and manual expert reviews of cases. In this study, a QA event was defined as any deviation identified through manual or automatic QA methods. It is important to note that the identification of a QA event did not necessarily indicate an error, but rather required further in- vestigation. QA events ranged from minor discrepancies to more severe issues, all of which were evaluated to ensure data and treatment quality.

Automatic data quality solutions

Automatic methods were defined as functions that run in the background and con– tinuously checked the database for inconsistencies. An overview of all methods can be found in Table 1. Algorithms were chosen based on experience from previous QA efforts [26], to cover areas of data reporting that were seen as susceptible to inconsistencies.

Methods were divided into two categories:

Pre-defined rules: Checks that worked based on pre-defined rules followed a clear decision tree. Thus there was no ambiguity in classifying each data point into correct and incorrect. Three important aspects with respect to the final analysis of the study data were surveyed with this approach: TNMT-Stage [27], EBRT elective targets and missing critical values. The first two methods cross-checked the selected value in the database with clinical and imaging information from patient status at diagnosis (see Figure B.7). The latter verified if all important parameters (see A.1) were available.

Machine learning/Statistics: On the other hand, issues where no clear decision boundary could be drawn were detected with statistical methods. These cases repre- sented uni- or multivariate

Table 1
Overview of automatic QA methods that are implemented in the tool. Techniques are divided into machine learning (ML)/statistics based- methods, and pre-defined rules. Number of true positives (TP) false negatives (FN) and false positive (FP) classifications are reported alongside precision and recall as performance metrics.

	Algorithm	TP[2]/FN[3]/FP[4]	Precision	Recall
	Machine learning/Statistics			
Outliers	Isolation Forest	99/25/2	99 / 101 (98 %)	99 / 124 (80 %)
IR-CTV volume consistency	Coefficient of Variation	7/0/0	7 / 7 (100 %)	7 / 7 (100 %)
DVH Relations	Mahalanobis Distance	130/43/52	130 / 182 (71 %)	130 / 173 (75 %)
	Pre-Defined Rules			
Missing Critical Value	Check for missing if treatment completed	-	157 / 157 (100 %)[1]	-
T-Stage consistency	Compare selected TNMT-Stage with diagnostic information	-	78 / 78 (100 %)[1]	-
EBRT Elective Target	Compare selected EBRT elective target with diagnostic information	-	121 / 121 (100 %)[1]	-

[1] Findings are based on unambiguous criteria.
[2] True positive.
[3] False negative.
[4] False positive.

anomalies that were impractical or impossible to capture with pre-defined rules.

Three different statistical algorithms were implemented:

1. First, a machine learning model was trained to automatically detect univariate outliers in all continuous variables of the database. This method aimed to catch data points that significantly differed from the norm and most likely originated from true deviations in the protocol or reporting errors."Isolation Forest", a decision tree-based anomaly detection algorithm was used for this purpose [28]. It is an unsupervised algorithm that works by creating multiple decision trees and partitioning the data set into smaller subsets. Intuitively, this outlier detection algorithm"isolates" observations by randomly selecting a split value between the maximum and minimum values of a selected feature. This recursive partitioning is repeated until all observations are isolated. For each observation, it returns an anomaly score based on the number of splits required to isolate a datapoint. The algorithm was chosen because of its computational efficiency, and the advantage that it works well when no anomalies are present in the training set. A separate model was constructed for each continuous variable with at least 100 non-empty values. The resulting models classified each data- point into normal and abnormal, based on an anomaly score threshold of 0.85.
2. The second method aimed to detect variations in the Intermediate Risk Clini- cal Target Volume (CTV-IR) volume, which is defined based on initial tumor extension at the time of diagnosis. Therefore, no significant variation across BT fractions would be expected. However, in clinical practice factors like inter-observer variations and imaging technique lead to non-zero differences in most patients. While some variation in reported CTV-IR volumes across BT frac- tions were therefore anticipated, flagging large implausible variations was an important aspect of BT QA.

The coefficient of variation (c_v), defined as the ratio of standard deviation and mean $c_v = \frac{\alpha}{\mu'}$, was used for this purpose. As a measure of relative variability it is widely used in statistics and data analysis, because it offers a simple and robust way to measure relative variability. A conservative threshold of $c_v > 0.33$ was chosen to flag implausible events.

3. Finally, some observations may only appear abnormal when two or more vari- ables were included in the analysis. In EMBRACE-II, discrete DVH parameters were collected to represent the dose distribution. The aim was to automatically flag inconsistencies in dose reporting, based on multivariate DVH parameter distributions, using Mahalanobis Distance *MD*.

MD was calculated as the distance of point $\vec{x}$ from the center of the distribution $\vec{\mu}\mu$ while taking into account covariance (Σ) between variables:

$$MD^2 = (\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu}) \quad (1)$$

Five 3D and two 2D data distributions of DVH parameters were investigated (Table A.2). Classification into normal and anomalous data points followed the process of:

a) Ensuring that the data were multivariate normally distributed by applying logarithmic transformation if necessary
b) Calculation of MD for each observation. Manual definition of a classifica- tion threshold in terms of upper quantile Q of the respective Chi-Square distribution with d degrees of freedom.
c) All samples with MD > Q were declared as an anomaly (Appendix Figure B.6)

MD thresholds Q were chosen conservatively based on observed variations of DVH distributions. For a list of all thresholds see Table A.2.

Manual review

In addition to automatic checks, all data was further manually reviewed by an ex- tended 10 members expert panel of the EMBRACE-II research team. During the review process researchers focused on their individual area of expertise and had ac- cess to the developed dashboard and study database. The expert panel reviewed 31/51 centers (1001/1475 patients). At least one physicist and radiation oncologist from the pool of 10 experts reviewed each case, Weekly discussions among the QA group were used to present findings and discuss questions with the whole panel of experts.

Evaluation

To assess the performance of the automatic detection algorithms, each finding was manually classified into either true positive, false negative or false positive by an expert panel of medical physicists and radiation oncologists. True positives were defined as a findings that, based on the observed values, a human reviewer would forward to the respective center for verification or clarification. Manually detected anomalies that went undetected by the automated methods were classified as false negatives. Precision and recall were calculated as performance metrics.

Furthermore, to learn about prevalent pitfalls in current MR-IGABT practice, all identified QA events were manually categorized. First, items were divided into events that arose due to missing or incomplete data, and events from other origins. Entries that were

not related to data completeness were analyzed in more detail. These were categorized into one of six classes, based on their origin with respect to the radiotherapy workflow. An overview of all classes, and their respective proportion of the total findings can be found in Table 2. In addition, they were also divided based on their severity, and impact on the study (Table 3).

Results

Fig. 2 shows several representative screenshots of the developed dashboard for an example center. The results of the survey are summarized in Fig. 3. The profession of survey participants was indicated as Radiation Oncologist (7/10) and Physicist (3/10). Results show generally favorable reactions to the prototype. All features were deemed to be valuable assets. Overall the platform was perceived as very helpful. Most users indicated that it is used for self-learning and departmental discussions. The number of automatically identified QA events for the different algo- rithms were; Isolation forest (IF): 101, coefficient of variation (CV): 7, Mahalanobis distance (MD): 182, missing critical value: 157, T-stage consistency: 78, and EBRT elective target: 121.

IF had a precision of 98 % and a recall of 80 %. Only 2/101 cases were deemed to be insignificant findings by human experts. One related to the reported body height of a patient, which was subsequently judged as plausible given its geographic origin and age. The other was related to very high, but possible creatinine clearance (Cockroft). CV had a precision and recall of 100 %, hence all identified cases with high variance in CTV-IR volumes were deemed to be implausible within known uncertainties, and no additional cases were flagged by manual review. MD had a precision of 71 % and a recall of 75 %. While no clear trend could be identified, many of the incorrectly flagged cases involved the placement of dosimetric reference points during treatment planning (ICRU-RV, ICRU Bladder, PIBS). Detailed values on True Positives, False Negatives and False Positives are shown in Table1.

Overall 2302 QA events were identified. 1235 (54 %) items were attributed to missing or incomplete data, and 1067 (46 %) originated from other sources. 646 (28 %) were found via automatic methods, and 1656 (72 %) by manual case reviews. At least one event, not related to data completeness, was found in 519/1001 (52 %) of patients. The number of patients with High Impact, and Low Impact events was 16 %, and 36 %, respectively (Table 3). The allocation of EBRT elective targets and TNMT- Staging were identified as major areas of uncertainty, with the former present in 12 %, and the latter in 8 % of patients (Table 2).

The number of events, not related to data completeness (Table 2), were related to *Treatment Technique* 287 (27 %), *Patient Characteristics* 235 (22 %), *Dose Reporting* 181 (17 %), *Outcome* 156 (15 %), *Outlier* 126 (12 %), and *RT Structures* 82 (8 %).

Discussion

Using novel software tools to monitor clinical trials in the pharmaceutical field has gained significant traction in the past decade.

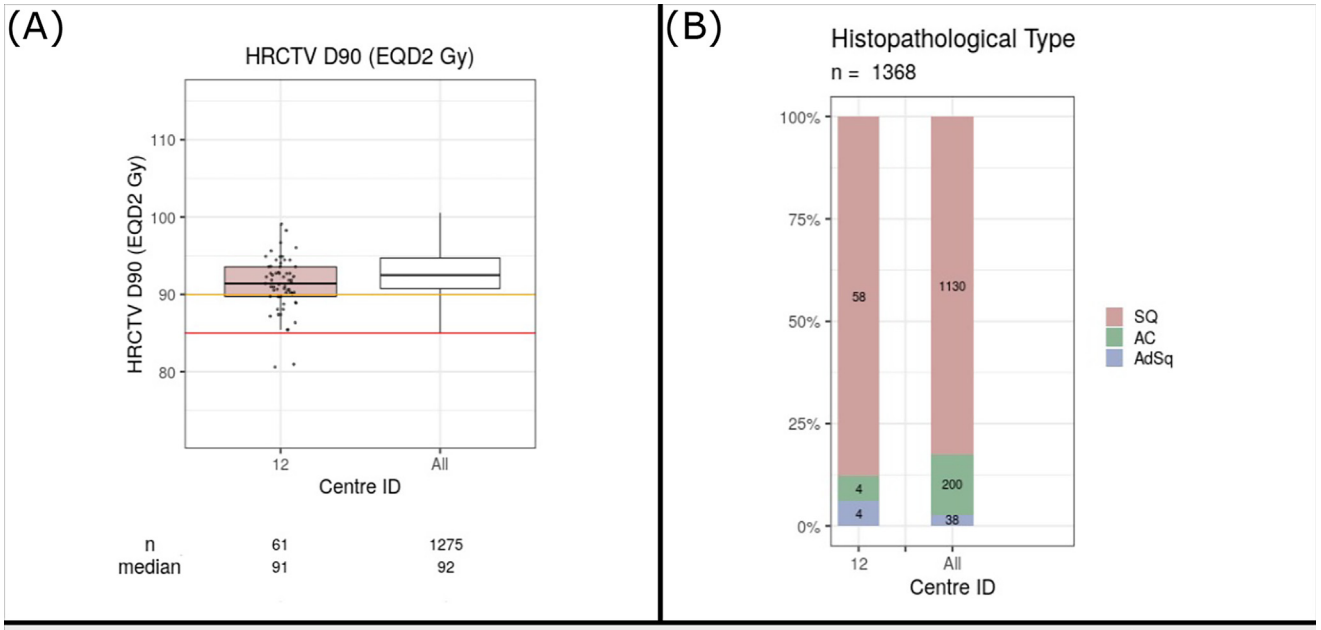
Table 2
Categorization of QA events, not related to data completeness, based on their origin in the treatment workflow. Proportion: number and percentage of all events. Last column shows the most common event, including its percentage within each group.

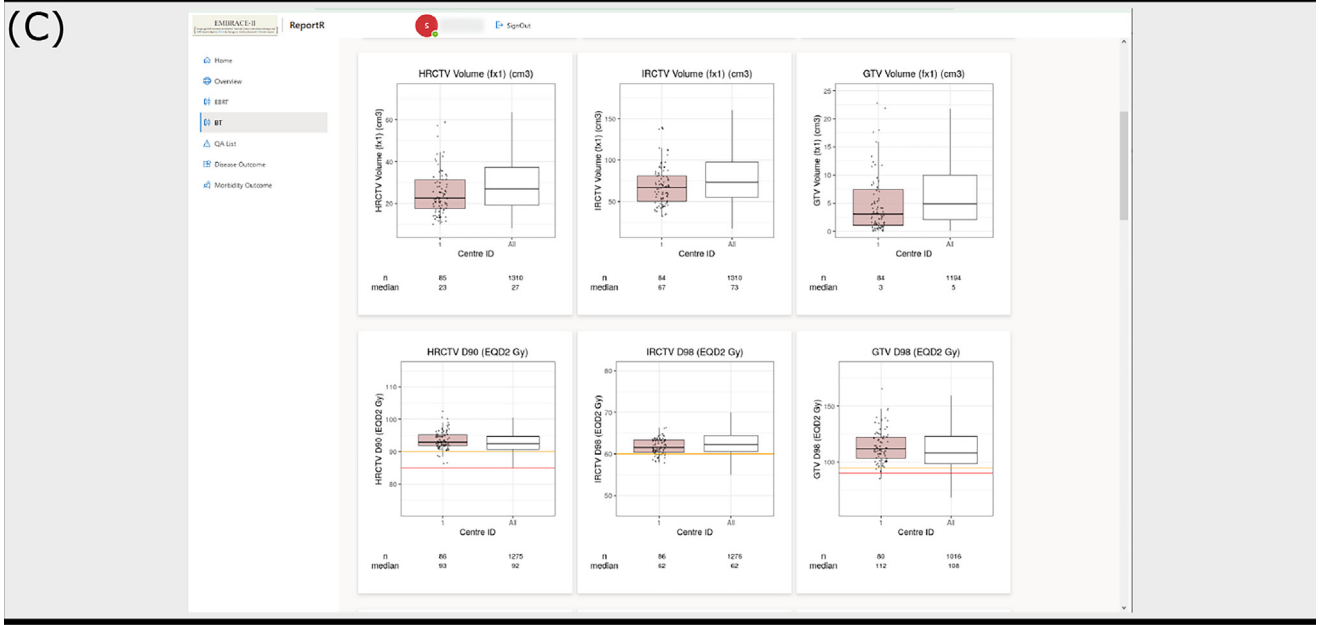
Origin	Description	Number[1]	Most Common Event[2]
Patient Characteristics	Patient and diagnostic related data such as TNM stage, OAR involvement, pathological node involvement or diagnostic procedures	235(22 %)/157/78	T-Stage consistency (34 %)
RT Structures	Errors related to characterization of structures used for dose prescription, e.g. contouring of targets and OAR, dose reference points or anatomical structures	82 (7.7 %)/29/53	Implausible relation between CTV-HR, CTV-IR and GTV volume: (10 %)
Treatment Technique	Factors influencing RT treatment procedure. Fractionation Schedule, EBRT elective targets, Usage of Needles/Applicator, Needle loading, TRAK, nodal boosting	287 (27 %)/156/131	EBRT Elective Target: (42 %)
Dose Reporting	DVH parameters that characterize the clinically used dose distribution. Dose optimization, violation of protocol limits for targets and OAR, implausible DVH relations	181 (17 %)/55/126	Implausible relation between ICRU-RV and Vagina reference points: (11 %)
Outcome	Patient status after treatment completion. Morbidity and Disease Events, Vital Status	156 (15 %)/156/0	Definition of local disease status: (13 %)
Outlier	Implausible values, transcription errors, significant deviation from the norm	126 (12 %)/25/101	FUP date reporting: (2 %)

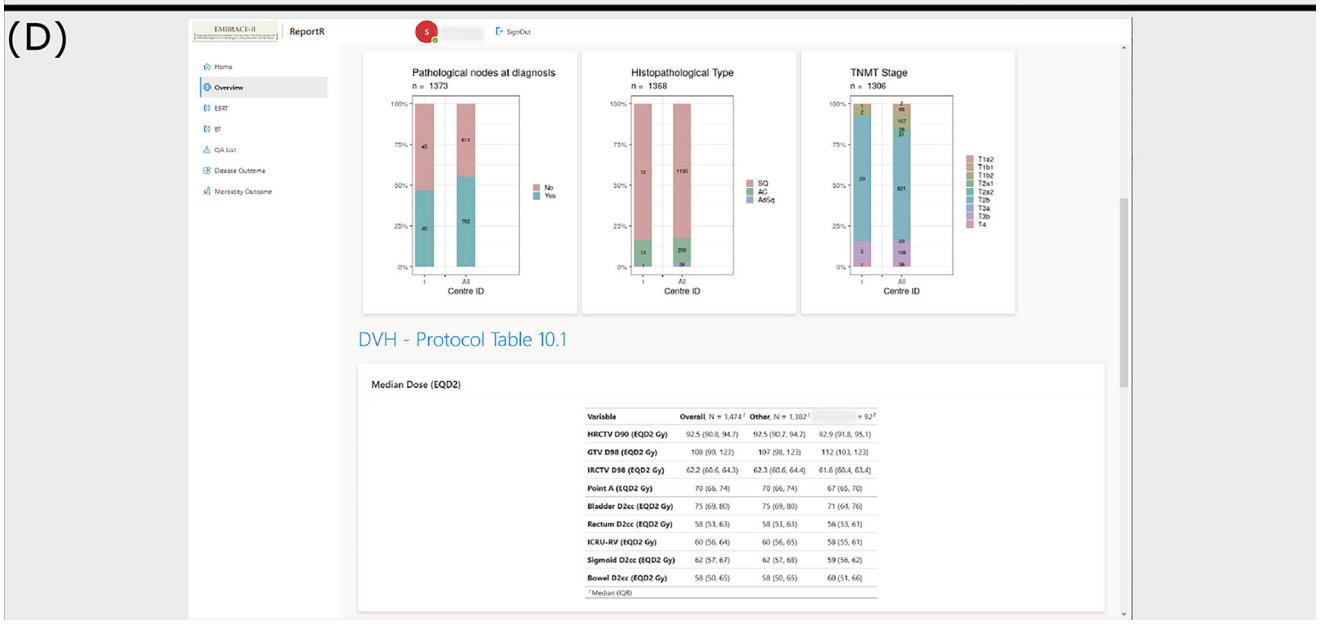
[1] n total (percent of all findings)/n manual/n automatic.
[2] (percent of category).

Table 3
Categorization of QA events, not related to data completeness, based on their severity and impact on the study.

Category	Definition	Examples of QA events	Number of patients (n = 1,001)
High Impact	Events that have high impact on study integrity and outcome analyses	Primary study endpoints Local control Nodal control Systemic control Vital status Morbidity Quality of life Violation of dose limits Tumor staging	162 (16 %)
Low Impact	Reporting errors Minor deviations from study protocol Variations of institutional practice Implausible data entries	Typing errors Inter-observer variations Dose optimization Exceeding overall treatment time EBRT elective targets	357 (36 %)
No Event	-	-	482 (48 %)


(A)
HRCTV D90 (EQD2 Gy)
HRCTV D90 (EQD2 Gy)
110
100
90
80
12
All
Centre ID
n 61 1275
median 91 92
(B)
Histopathological Type
n = 1368
100%
75%
50%
25%
0%
58
4
4
1130
200
38
12
All
Centre ID
SQ
AC
AdSq


(C)
EMBRACE-II
ReportR
SignOut
Home
Overview
EBRT
BT
QA List
Disease Outcome
Morbidity Outcome
HRCTV Volume (fx1) (cm3)
IRCTV Volume (fx1) (cm3)
GTV Volume (fx1) (cm3)
HRCTV D90 (EQD2 Gy)
IRCTV D98 (EQD2 Gy)
GTV D98 (EQD2 Gy)
Centre ID


(D)
EMBRACE-II
ReportR
SignOut
Home
Overview
EBRT
BT
QA List
Disease Outcome
Morbidity Outcome
Pathological nodes at diagnosis
n = 1373
Histopathological Type
n = 1368
TNMT Stage
n = 1306
Centre ID
DVH - Protocol Table 10.1
Median Dose (EQD2)
Variable
HRCTV D90 (EQD2 Gy)
GTV D98 (EQD2 Gy)
IRCTV D98 (EQD2 Gy)
Point A (EQD2 Gy)
Bladder D2cc (EQD2 Gy)
Rectum D2cc (EQD2 Gy)
ICRU-RV (EQD2 Gy)
Sigmoid D2cc (EQD2 Gy)
Bowel D2cc (EQD2 Gy)

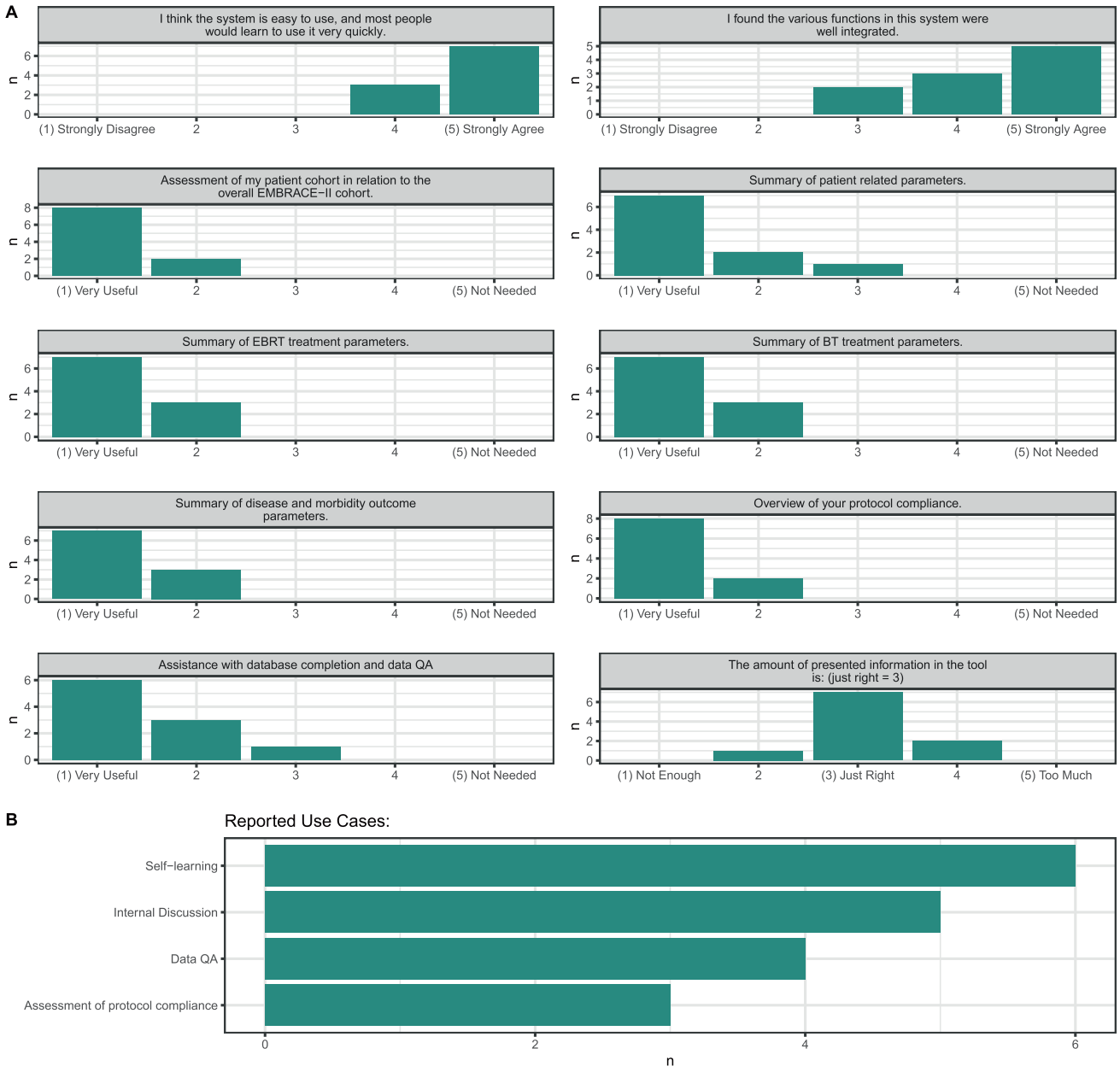


Fig. 3. Results of the anonymous survey among 10 participating centers. Panel (A) shows responses regarding utility and usability of the tool. Panel (B) summarizes for what purpose the tool was used among re- sponders. In this case multiple answers were allowed.

Driven by economic considerations, centralized monitoring and statistical monitoring are now being recommended by both the FDA and EMA to effectively handle the ever increasing volume and va- riety of collected data [29–32]. However, to the best of our knowledge, no comparable projects exist in radiation oncology. Using modern IT infrastructure facilitated secure access to individ- ualized dashboards for multiple authorized users. Feedback from the small survey indicated good usability, and potential value of the tools features. The individualized dashboards indicated that even among advanced EMBRACE-II centers, there can be consider- able differences in treatment approach. While this outcome was expected and is acknowledged in the community [26,33], the developed project offered additional evidence, but also quantifi- able metrics re- garding this topic. It could serve as a platform to monitor and study inter-center uncertainties in the future, and assist in implementing a high standard of care for LACC patients treated with EBRT and MR-IGABT. However, it should be noted that the tool's ability to compare individual institutional/patient data with a refer- ence cohort may be limited by the amount of avail- able data. In cases of smaller trials or institutions with limited patient data, data from other sources may need to be utilized to establish a larger reference set. Beyond interventional clinical trials

Fig. 2. Example screenshots of the dashboard. Panel (A) shows a represen- tative figure that summarizes continuous variables with boxplots for the center (left) and the overall cohort (right). Shown here is the dose to the HRCTV D90 in EQD2 Gy. Red and orange horizontal lines represent protocol dose limits and planning aims, respectively. In analogue, panel (B) shows a summary graph for categorical vari- ables, in this example histopathological type at diagnosis. Panel (C) shows the entire interface of the application, with a navigation bar on the left, and various summary figures for treatment parameters in the center. Panel (D) shows several patient related parameters at the top, and a summary table of protocol compliance with respect to total dose limits to targets and organs at risk.

such as EMBRACE-II, the presented methodologies could also be adapted for routine data collection in radiation oncology, or for the set-up of registries.

Of the automatic QA methods, both isolation forest and coefficient of variation showed excellent precision. Given that the thresholds were chosen rather conserva- tively in order to prevent overwhelming detection of false positives, this result was in line with expectations. It is also acknowledged that some anomalies likely evaded manual detection, as this was still a human process, and that classification into"nor- mal" and"abnormal" can be borderline and subject to interpretation. These aspects should always be considered when interpreting the reported numbers.

Nevertheless, the use of Isolation Forest to automatically scan univariate data for outliers proved to be a computationally inexpensive and flexible method. Most out- liers could be attributed to human transcription errors in the eCRF, underlying that for reporting and documentation of RT data, eliminating the human factor to prevent inaccuracies would be of considerable interest.

The results of using the coefficient of variation to detect variations in the CTV-IR volume suggest that this method is effective in identifying substantial deviations across fractions. However, the threshold was set to specifically detect only major variations, as inter-observer variations in radiation oncology are known to be sub- stantial [34]. For example, a study by Petric et al. report a volumetric conformity index for expert consensus contours of only 0.68 for CTV-IR [35]. As a result, while this method may be suitable for detecting significant implausible variations, it may not be as sensitive to more subtle deviations and can only be applied to parameters that are expected to remain constant throughout treatment. Although only a limited number of implausible CTV-IR volume variations were detected, none were identified through human review. Again, this is likely due to the reviewers' expectation of large inter-observer variations. Nonetheless, the identified cases are important to highlight as they likely resulted from other sources of error.

Detection of implausible reported treatment plans through multivariate analysis of DVH parameters, proved to be a more challenging task. The comparatively low precision of 71 % highlighted that more sophisticated methods may be required to automate this task. For example, Li et al. showed how Gower distance can be used to flag anomalous prescriptions in radiation oncology[36]. As can be seen from the comparatively large number of false positives and false negatives, additional infor- mation from diagnostic and treatment variables may be required to model existing variations in the data. Nevertheless, it was demonstrated that the method could identify many reporting inconsistencies, based on discrete DVH metrics alone.

Overall, it can be seen that automatic anomaly detection methods have the potential to play a valuable role in clinical trial QA and in clinical practice for radiation oncology. As demonstrated in this work, even simple methods can help to identify and flag unusual data points that deviate from expected patterns, which could serve as an additional safety measure. Examples of how such methods could be used in radiation oncology include:

- Treatment plans: Identify unusual treatment plans e.g., if a treatment plan has an unexpected high dose to OAR or an implausible target volume.
- Treatment delivery: Check for proper machine and patient setup using treat- ment machine parameters (e.g. log files), and identify potential issues early on.
- Patient outcome: Monitor toxicity and survival rates, and alert trial coordina- tors.

However, as shown in this work, finding a suitable threshold for anomaly detection can be difficult, as it depends on data characteristics, and the domain of it's appli- cation. Thus, it is important to note that, while these automatic methods can help to ensure the accuracy and integrity of the data, they should be used with manual review and oversight by trained medical professionals.

While the tool cannot go into as much depth as manual expert review of individual patients, it enabled researchers and centers to identify patterns in reported data, sub- stantially save human effort and flag deviations more efficiently. For 1001 patients, 2302 QA events were identified. This would correspond to 0.4 % of all collected data (assuming 600 data points per patient). However, reporting of a robust number is challenging as the number of mandatory fields in the database varied among patients, and analysing each case in full detail was not feasible. Therefore the reported event rate should be interpreted as a rough estimate. While this indicates that overall study data was of high quality, in 16 % of patients at least one high impact event was detected, which highlights the value of thorough QA efforts.

Overall, issues regarding treatment technique were most common. This could orig- inate from the complexity of the treatment, that offered a wide range of treatment options for EBRT and BT. However, since the analysis was based on retrospective data collected in a clinical study, the relations between defined QA origins may not be directly representative for the clinical environment. As no DICOM RT files were directly collected, treatment plans could only be evaluated based on reported DVH parameters. This limited the assessment of structure definitions and dose distribu- tions, and likely resulted in underestimation of these events in comparison to clinical practice. The frequent inconsistencies in allocation of EBRT elective targets and TNMT-Staging may require future investigation. While TNM staging uncertainties are a known phenomenon [37], the elective target concept is a newly introduced con– cept in EMBRACE-II, which intends to adapt the EBRT target volume to patient- specific risk factors. Novelty of the concept may explain frequent inconsistencies, however deeper analysis would be warranted.

The analysis of QA events was based on a list that was generated through automatic checks and manual expert case reviews. Both methods were inherently biased. Au- tomatic checks that were implemented were chosen due to pre-existing knowledge about pitfalls [26] and cover only a small fraction of recorded data. Likewise, human experts reviewers tended to focus on their field of expertise. This bias was reflected in the reality that some issues gained more attention than others. Collection of structured, high-quality data was a prerequisite for this project. In this case such a database was already established due to the underlying EMBRACE-II study. It is acknowledged that for translation into clinical practice, such routine data collection would require overcoming several significant practical challenges at first.

Nevertheless, we believe that our work can serve as an example of what could be achieved once these barriers are overcome. We believe that our results show that it would be of considerable interest to extend automatic QA algorithms, as presented in this work, beyond applications in clinical trials.

Conclusion

A software was developed and tested within a clinical trial in radiation oncology for cervical cancer. This tool enabled the quantitative and qualitative comparison of institutional patient and treatment characteristics and outcome data, with a large multicenter reference patient cohort. We demonstrated the value of using statisti- cal methods to automatically detect implausible data points, and highlighted com- mon pitfalls and uncertainties in radiotherapy for cervical cancer. A comprehensive framework was presented that could serve as a blueprint for advanced data

analysis methods for clinical studies in radiation oncology, and beyond.

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.

Acknowledgements

This research was funded by the Austrian Science Fund (FWF, project number KLI695-B33), Austria.

The code was written using R version 4.0.2 using open source software. The code may be made available upon request, from the corresponding author.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2023.109524.

References

[1] Fiorino C, Jeraj R, Clark CH, Garibaldi C, Georg D, Muren L, et al. Grand challenges for medical physics in radiation oncology. Radiother Oncol Dec. 2020;153:7–14.
[2] Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med Jan. 2022;28:31–8.
[3] O. Morin, M. Valli'eres, S. Braunstein, J. B. Ginart, T. Upadhaya, H. C. Woodruff, A. Zwanenburg, A. Chatterjee, J. E. Villanueva-Meyer, G. Valdes, W. Chen, J. C. Hong, S. S. Yom, T. D. Solberg, S. Ľock, J. Seuntjens, C. Park, and P. Lambin, "An artificial intelligence framework integrating longitudinal electronic health records with real-world data enables continuous pan-cancer prognostication," *Nature Cancer*, vol. 2, pp. 709–722, July 2021.
[4] Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis May 2020;20:533–4.
[5] Sahiner B, Pezeshk A, Hadjiiski LM, Wang X, Drukker K, Cha KH, et al. Deep learning in medical imaging and radiation therapy. Med Phys Jan. 2019;46:e1–e36.
[6] Lambin P, van Stiphout RGPM, Starmans MHW, Rios-Velazquez E, Nalbantov G, Aerts HJWL, et al. Predict- ing outcomes in radiation oncology—multifactorial decision support systems. Nat Rev Clin Oncol Jan. 2013;10:27–40.
[7] McNutt TR, Moore KL, Wu B, Wright JL. Use of big data for quality assurance in radiation therapy. Semin Radiat Oncol Oct. 2019;29:326–32.
[8] Chan MF, Witztum A, Valdes G. Integration of AI and machine learn- ing in radiotherapy QA. Front Artif Intell Sept. 2020;3:577620.
[9] Kalet AM, Luk SMH, Phillips MH. Radiation therapy quality assurance tasks and tools: the many roles of machine learning. Med Phys May 2020;47.
[10] Cohen PA, Jhingran A, Oaknin A, Denny L. Cervical cancer. Lancet Jan. 2019;393:169–82.
[11] Rodin D, Burger EA, Atun R, Barton M, Gospodarowicz M, Grover S, et al. Scale-up of radiotherapy for cervical cancer in the era of human papillomavirus vaccination in low-income and middle-income countries: a model-based analysis of need and economic impact. Lancet Oncol July 2019;20:915–23.
[12] D. Cibula, R. Ṕotter, F. Planchamp, E. Avall-Lundqvist, D. Fischerova, C. Haie-Meder, C. Ḱohler, F. Landoni, S. Lax, J. C. Lindegaard, U. Mahantshetty, P. Mathevet, W. G. McCluggage, M. McCormack, R. Naik, R. Nout, S. Pignata, J. Ponce, D. Querleu, F. Raspagliesi, A. Rodolakis, K. Tamussino, P. Wim- berger, and M. R. Raspollini, "The European Society of Gynaecological On- cology/European Society for Radiotherapy and Oncology/European Society of Pathology Guidelines for the Management of Patients with Cervical Cancer," *Virchows Archiv*, vol. 472, pp. 919–936, June 2018.
[13] Sturdza AE, Knoth J. Image-guided brachytherapy in cervical cancer including fractionation. Int J Gynecol Cancer Mar. 2022;32:273–80.
[14] R. Ṕotter, K. Tanderup, C. Kirisits, A. de Leeuw, K. Kirchheiner, R. Nout, L. T. Tan, C. Haie-Meder, U. Mahantshetty, B. Segedin, P. Hoskin, K. Bruheim, B. Rai, F. Huang, E. Van Limbergen, M. Schmid, N. Nesvacil, A. Sturdza, L. Fokdal, N. B. K. Jensen, D. Georg, M. Assenholt, Y. Seppenwoolde, C. Nom- den, I. Fortin, S. Chopra, U. van der Heide, T. Rumpold, J. C. Lindegaard, and I. Jürgenliemk-Schulz, "The EMBRACE II study: The outcome and prospect of two decades of evolution within the GEC-ESTRO GYN working group and the EMBRACE studies," *Clinical and Translational Radiation Oncology*, vol. 9, pp. 48–60, Feb. 2018
[15] "International Commission on Radiation Units and Measurements," *Journal of the ICRU*, vol. 13, pp. NP.2–NP, Apr. 2013.
[16] Viswanathan AN, Creutzberg CL, Craighead P, McCormack M, Toita T, Narayan K, et al. International brachytherapy practice pat- terns: a survey of the gynecologic cancer intergroup (GCIG). Int J Radiat Oncol Biol Phys Jan. 2012;82:250–5.
[17] R. Ṕotter, C. Haie-Meder, E. V. Limbergen, I. Barillot, M. D. Brabandere, J. Dimopoulos, I. Dumas, B. Erickson, S. Lang, A. Nulens, P. Petrow, J. Rownd, and C. Kirisits, "Recommendations from gynaecological (GYN) GEC ESTRO working group (II): Concepts and terms in 3D image-based treatment planning in cervix cancer brachytherapy—3D dose volume parameters and aspects of 3D image-based anatomy, radiation physics, radiobiology," *Radiotherapy and Oncology*, vol. 78, pp. 67–77, Jan. 2006
[18] C. Haie-Meder, R. Ṕotter, E. Van Limbergen, E. Briot, M. De Brabandere, J. Dimopoulos, I. Dumas, T. P. Hellebust, C. Kirisits, S. Lang, S. Muschitz, J. Nevinson, A. Nulens, P. Petrow, and N. Wachter-Gerstner, "Recommenda- tions from Gynaecological (GYN) GEC-ESTRO Working Group (I): concepts and terms in 3D image based 3D treatment planning in cervix cancer brachytherapy with emphasis on MRI assessment of GTV and CTV," *Radiotherapy and Oncology*, vol. 74, pp. 235–245, Mar. 2005
[19] T. P. Hellebust, C. Kirisits, D. Berger, J. Ṕerez-Calatayud, M. De Braban- dere, A. De Leeuw, I. Dumas, R. Hudej, G. Lowe, R. Wills, and K. Tanderup, "Recommendations from Gynaecological (GYN) GEC-ESTRO Working Group: Considerations and pitfalls in commissioning and applicator reconstruction in 3D image-based treatment planning of cervix cancer brachytherapy," *Radiother- apy and Oncology*, vol. 96, pp. 153–160, Aug. 2010.
[20] Dimopoulos JC, Petrow P, Tanderup K, Petric P, Berger D, Kirisits C, et al. Recommen- dations from Gynaecological (GYN) GEC-ESTRO Working Group (IV): basic principles and parameters for MR imaging within the frame of image based adaptive cervix cancer brachytherapy. Radiother Oncol Apr. 2012;103:113–22.
[21] R. Ṕotter, K. Tanderup, M. P. Schmid, I. Jürgenliemk-Schulz, C. Haie-Meder, L. U. Fokdal, A. E. Sturdza, P. Hoskin, U. Mahantshetty, B. Segedin, K. Bruheim, F. Huang, B. Rai, R. Cooper, E. van der Steen-Banasik, E. Van Limbergen, B. R. Pieters, L.-T. Tan, R. A. Nout, A. A. C. De Leeuw, R. Ristl, P. Petric, N. Nesvacil, K. Kirchheiner, C. Kirisits, J. C. Lindegaard, C. Chargari, I. Dumas, G. Lowe, J. Swamidas, R. Hudej, T. Paulsen Helle- bust, G. Menon, A. S. Oinam, P. Bownes, M. Christiaens, M. De Braban- dere, H. Janssen, B. Oosterveld, K. Koedooder, A. B. Langeland Marthinsen, M. Sundset, D. Whitney, M. Ketelaars, L. C. Lutgens, B. Reinniers, I. Mora, E. Villafranca, G. Antal, J. Hadjiev, F. Bachand, D. Batchelar, B. Erickson, J. Rownd, G. Jacobson, Y. Kim, M. Anttila, J.-E. Palmgren, J. An, M. S. Assenholt, S. Banerjee, S. Bentzen, T. Berger, P. Dankulchai, T. Diendor- fer, I. Dilworth, J. Dimopoulos, E. Ďorr, S. Ecker, M. Federico, E. Fidarova, I. Fortin, P. Georg, J. Gora, N. Hegazy, N. Jastaniyah, N. B. K. Jensen, T. Liederer, K. Majercakova, D. Misimovic, L. Motisi, D. Najjari Jamal, K. Nki- wane, A. Schwartz-Vittrup, M. Serban, S. Smet, S. Spampinato, P. Trnkova, M. Valgma, H. Westerveld, J. S. Y. Wong, and K. Yoshida, "MRI-guided adap- tive brachytherapy in locally advanced cervical cancer (EMBRACE-I): a multicentre prospective cohort study," *The Lancet Oncology*, vol. 22, pp. 538–547, Apr. 2021
[22] Han K, Milosevic M, Fyles A, Pintilie M, Viswanathan AN. Trends in the utilization of brachytherapy in cervical cancer in the United States. Int J Radiat Oncol Biol Phys Sept. 2013;87:111–9.
[23] L.-T. Tan, K. Tanderup, C. Kirisits, U. Mahantshetty, J. Swamidas, I. Jürgenliemk-Schulz, J. Lindegaard, A. de Leeuw, N. Nesvacil, M. Assen- holt, D. Berger, T. Diendorfer, J. Dimopoulos, S. Duke, S. Ecker, L. Fokdal, T. Hellebust, N. Jensen, K. Kirchheiner, R. Nout, P. Petric, M. Schmid, Y. Sep- penwoolde, A. Sturdza, E. Van Limbergen, C. Haie-Meder, and R. Ṕotter, "Education and training for image-guided adaptive brachytherapy for cervix cancer—The (GEC)-ESTRO/EMBRACE perspective," *Brachytherapy*, vol. 19, pp. 827–836, Nov. 2020
[24] W. Chang, J. Cheng, J. Allaire, C. Sievert, B. Schloerke, Y. Xie, J. Allen, J. McPherson, A. Dipert, and B. Borges, *shiny: Web Application Framework for R*, 2021. R package version 1.7.1
[25] J. Brooke, "SUS – a quick and dirty usability scale," pp. 189–194, Jan. 1996.
[26] Kirisits C, Federico M, Nkiwane K, Fidarova E, Jürgenliemk-Schulz I, de Leeuw A, Lindegaard J, Ṕotter R, Tanderup K. Quality assurance in MR image guided adaptive brachytherapy for cervical cancer: final results of the EMBRACE study dummy run. Radiother Oncol Dec. 2015;117:548–54.
[27] Brierley JD, Gospodarowicz MK, Wittekind C. TNM classification of malignant tumours. John Wiley & Sons; 2017.
[28] F. T. Liu, K. Ting, and Z.-H. Zhou, "Isolation Forest," pp. 413 – 422, Jan. 2009.
[29] F. D. A. (FDA), "Oversight of Clinical Investigations — A Risk-Based Approach to Monitoring," p. 22.
[30] E. M. A. (EMA), "Reflection paper on risk based quality management in clinical trials,".
[31] Olsen R, Bihlet AR, Kalakou F, Andersen JR. The impact of clinical trial monitoring approaches on data integrity and cost—a review of current literature. Eur J Clin Pharmacol Apr. 2016;72:399–412.
[32] O. T. Inan, P. Tenaerts, S. A. Prindiville, H. R. Reynolds, D. S. Dizon, K. Cooper-Arnold, M. Turakhia, M. J. Pletcher, K. L. Preston, H. M. Krumholz, B. M. Marlin, K. D. Mandl, P. Klasnja, B. Spring, E. Iturriaga, R. Campo, P. Desvigne-Nickens, Y. Rosenberg, S. R. Steinhubl, and R. M. Califf, "Digitiz- ing clinical trials," *npj Digital Medicine*, vol. 3, p. 101, Dec. 2020.
[33] Viswanathan AN, Erickson B, Gaffney DK, Beriwal S, Bhatia SK, Lee Burnett O, et al. Comparison and consensus guidelines for de- lineation of clinical target volume for CT- and MR-based brachytherapy in locally advanced cervical cancer. Int J Radiat Oncol Biol Phys Oct. 2014;90:320–8.
[34] Vinod SK, Jameson MG, Min M, Holloway LC. Uncertainties in volume delineation in radiation oncology: a systematic review and recommendations for future studies. Radiother Oncol 2016;121:169–79.

[35] Petrič P, Hudej R, Rogelj P, Blas M, Tanderup K, Fidarova E, et al. Uncertainties of target volume delineation in MRI guided adaptive brachytherapy of cervix can- cer: a multi-institutional study. Radiother Oncol Apr. 2013;107:6–12.

[36] Q. Li, J. Wright, R. Hales, R. Voong, and T. McNutt, "A digital physician peer to automatically detect erroneous prescriptions in radiotherapy," *npj Digital Medicine*, vol. 5, p. 158, Oct. 2022.

[37] J. Knoth, R. Ṗotter, I. Jürgenliemk-Schulz, C. Haie-Meder, L. Fokdal, A. Stur- dza, P. Hoskin, U. Mahantshetty, B. Segedin, K. Bruheim, E. Wiebe, B. Rai,R. Cooper, E. van der Steen-Banasik, E. van Limbergen, B. Pieters, M. Sund- set, L. Tan, R. Nout, K. Tanderup, C. Kirisits, N. Nesvacil, J. Lindegaard, and M. Schmid, "Clinical and imaging findings in cervical cancer and their impact on FIGO and TNM staging – An analysis from the EMBRACE study," *Gynecologic Oncology*, vol. 159, pp. 136–141, Oct. 2020