External Validation of a Bayesian Network for Error Detection in Radiotherapy Plans

Petros Kalendralis, Denis Eyssen, Richard Canters, Samuel M. H. Luk, Alan M. Kalet, Wouter van Elmpt, Rianne Fijten, Andre Dekker, Catharina M. L. Zegers, and Inigo Bermejo

***Abstract*—Artificial intelligence (AI) applications have recently been proposed to detect errors in radiotherapy plans. External validation of such systems is essential to assess their performance and safety before applying them to clinical practice. We collected data from 5238 patients treated at Maastro Clinic and introduced a range of common radiotherapy plan errors for the model to detect. We estimated the model's discrimination by calculating the area under the receiver-operating characteristic curve (AUC). We also assessed its clinical usefulness as an alert system that could reduce the need for manual checks by calculating the percentage of values flagged as errors and the positive predictive value (PPV) for a range of high sensitivities (95%–99%) and error prevalence. The AUC when considering all variables was 67.8% (95% CI, 65.6%–69.9%). The AUC varied widely for different types of errors (from 90.4% for table angle errors to 54.5% for planning tumor volume-PTV dose errors). The percentage of flagged values ranged from 84% to 90% for sensitivities between 95% and 99% and the PPV was only slightly higher than the prevalence of the errors. The model's performance in the external validation was significantly worse than that in its original setting (AUC of 68% versus 89%). Its usefulness as an alert system to reduce the need for manual checks is questionable due to the low PPV and high percentage of values flagged as potential errors to achieve a high sensitivity. We analyzed the apparent limitations of the model and we proposed actions to overcome them.**

***Index Terms*—Artificial intelligence (AI), Bayesian network (BN), radiotherapy, treatment planning.**

I. Introduction

OVER the past decades, radiotherapy has constituted a fundamental treatment modality for cancer patients along with other treatment options, such as surgery, chemotherapy, and immunotherapy [1]. Radiotherapy's cost effectiveness (5% of the total cost of oncological care) [2] as well as the number of patients that are treated with it (50% of cancer patients) [1], [3] and its potentially curative nature [4], stress the need for an accurate treatment plan construction and delivery. Recent advancements in the field of artificial intelligence (AI) have contributed to a significant progress regarding the automation of the treatment planning process such as the automatic delineation of the clinical target volumes (CTVs) or organs at risk (OAR) [5], [6] and the automatic dosimetric evaluation of treatment planning [7].

Radiotherapy treatment planning is a complex procedure that requires a coordinated team effort by an interdisciplinary group that consists of radiation oncologists, medical physicists, radiation technologists, and dosimetrists. The objective of radiotherapy treatment planning is to safely and efficiently prescribe the optimal dose to the anatomical target volume of the patients. Mistakes made during this process can cause serious risks during the treatment planning execution. In the past, several organizations, such as the World Health Organization (WHO), the American Association of Physicists in Medicine (AAPM), and the European Society for Therapeutic Radiation Oncology (ESTRO), have published recommendation guidelines for the elimination of the radiotherapy errors [8]–[10]. Generally, the radiotherapy treatment plan errors can be subdivided into operational or system errors. For instance, malfunction of the multileaf collimators (MLCs) system of the linear accelerator (LINAC) in a case of intensity-modulated radiation therapy (IMRT) or differences between the prescribed dose and the dose per radiotherapy fraction due to adjustments of the reference points are some of the potential errors. These errors can lead to serious accidents with extremely severe consequences for both patients and clinical professionals [11], [12].

Increased automation, supported by AI techniques and combined with human expertise, could reduce the time needed for the development and execution of a radiotherapy treatment plan. Furthermore, the implementation of AI methods can potentially contribute to the early detection of plan errors and the reduction of the time needed for their detection [13], [14].

Currently, we are entering a new challenging and promising era in radiotherapy where AI has started to manifest its potential with several applications. For example, several studies introduced automated treatment plan verification for the detection or errors during radiotherapy [15]–[19]. Moreover, with the development of the automated pipelines for the validation and quality assurance (QA) of the radiotherapy plans, objections raised regarding their accuracy and implementation, such as the requirement of expertise knowledge of the

Manuscript received December 4, 2020; revised February 2, 2021 and March 26, 2021; accepted March 29, 2021. Date of publication April 2, 2021; date of current version February 3, 2022. *(Catharina M. L. Zegers and Inigo Bermejo contributed equally to this work.) (Corresponding author: Petros Kalendralis.)*

Petros Kalendralis, Denis Eyssen, Richard Canters, Wouter van Elmpt, Rianne Fijten, Andre Dekker, Catharina M. L. Zegers, and Inigo Bermejo are with the Department of Radiation Oncology (Maastro), GROW School for Oncology, Maastricht University Medical Centre+, 6229 ET Maastricht, The Netherlands (e-mail: petros.kalendralis@maastro.nl).

Samuel M. H. Luk and Alan M. Kalet are with the Department of Radiation Oncology, University of Washington Medical Center, Seattle, WA 98195 USA.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TRPMS.2021.3070656.

Digital Object Identifier 10.1109/TRPMS.2021.3070656

manual planning (i.e., human intervention) and reproducibility issues [20].

To address these limitations, Luk *et al.* [21] proposed a model to detect radiotherapy errors using an AI-based approach. Their Bayesian network (BN) model can flag anomalies in 29 variables related to diagnostic, prescription, plan, and setup level parameters to assist clinical physicists and clinicians on the time-consuming and error-prone radiotherapy treatment planning procedure.

BNs are the most popular type of probabilistic graphical models (PGMs), which emerged during the 1980s and rose to prominence in the next decade [22]. PGMs use graphs to represent the probabilistic dependencies between the variables in a model. BNs, for example, use directed acyclic graphs (DAGs) where each variable is represented by a node and links between variables imply causality. In addition, the conditional probability distribution (CPD) of each variable is defined as a function of its parents in the graph (i.e., the set of nodes that have links pointing at one particular node). The structure of the graph of the BN and the CPDs can be either defined based on expert knowledge or learned from data using machine learning algorithms [23]. Probabilistic reasoning in BNs allows for different types of queries, such as the probability distribution of one or more target variables given a set of findings (e.g., what is the probability of rain given the grass is wet), or the probability of a set of findings (e.g., what is the probability of rain and dry grass). A set of such findings is referred to as evidence. The intuitiveness of the probabilistic reasoning in BNs thanks to their graphical structure in contrast to black-box algorithms prominent in AI has led to a wide adoption in healthcare [24].

Luk *et al.* [21] defined the DAG based on expert knowledge and learned the CPDs based on historical data from their institution. Consequently, they showed that they could detect anomalies in radiotherapy plans assigning the values of a given radiotherapy plan to the variables of the BN and calculating the probability of the evidence, because radiotherapy plans with errors will generally result in a lower probability.

We hypothesized that such a model is clinically relevant and can provide significant added value, reducing the need for manual checks and detecting errors that would otherwise go unnoticed. An external validation is an empirical evaluation in a dataset that was not used to develop the model and they are essential before considering whether to use a clinical prediction model [25]. Therefore, we performed an external validation of the model using data from Maastro clinic (The Netherlands), with the aim to assess the generalizability of the model.

II. Materials and Methods

A. Data Acquisition

We used data from 5238 patients (19 054 treatment plans) for this study, collected at the Maastro radiation oncology clinic (Maastricht, The Netherlands) between 2012 and 2020. The patients were treated with external beam radiotherapy using electrons or photons with IMRT and volumetric-modulated arc therapy (VMAT) in seven different Truebeam LINACs of Varian medical systems. Patients treated with protons were excluded from the dataset as the original model by Luk *et al.* [21] did not include them. The radiotherapy elements were extracted and collected from the Varian

TABLE I
Description and Examples of the Maastro Clinic's Variables Used

Diagnostic variables			
Variable name	**Description**	**States number**	**Examples**
Diagnose	Anatomic tumor location	226	”prostaat”, ”long”
cT	Clinical T stage	29	0,1,1a
cN	Clinical N stage	17	1b,1c,2a
cM	Clinical M stage	8	1,1a,1b
Prescription variables			
Variable name	**Description**	**States number**	**Examples**
Treatment_ Intent	Treatment Intent	3	”Radicaal”, ”Palliatief”
NumberOf Rxs	Number of prescriptions	14	1,2,3
DosePer Fraction	Dose per fraction	61	2.75,2,4
PTVDoseRx	Total dose	210	15,20,48
Total Fractions	Fractions	35	4,5,8
RxRadiation Type	Radiation Type	7	6X,10X
Plan/Beam variables			
Variable name	**Description**	**States number**	**Examples**
Plan Technique	Planning technique	5	”ARC”, ”STATIC”
TableAngle	Table angle	50	0,10,355
NumberOf Beams	Number of beams	14	3,4,5
Wedge	Wedge position	1	0,00%
ControlPoints	Control points	714	6,8,10
SSD	Source to surface distance	15478	70.1,72.2,73.6
Bolus	Presence/ type of bolus	4	N,Y, ”MULTI-VALUE”
GantryAngle	Gantry angle	1329	103.8,104,104.1
Collimator Angle	Collimator angle	725	347.5,348.3, 354.9
BeamEnergy	Beam energy	6	2,3,4

Continued

TABLE I
(Continued.) DESCRIPTION AND EXAMPLES OF THE MAASTRO CLINIC'S VARIABLES USED

Setup variables			
Variable name	**Description**	**States number**	**Examples**
Orientation	Patient scan orientation	2	"Head First-Supine", "HeadFirst-Decubitus Right"
CouchLat	Lateral couch position	5681	17.1,326.3,-1.70
CouchLong	Longitudinal couch position	5743	103.4,108.7, 112.9
CouchVert	Vertical couch position	5707	-10.5, -5.4
Tolerance	Setup tolerance table	1	"Console RUIM"

TABLE II
ERRORS SIMULATION OVERVIEW

Errors category	**Errors description**	**Errors specification**
Patient positioning	Table rotation errors	Table rotation values bigger than 10 degrees from the planned value
Prescription level	Prescribed dose to the PTV is not equal to the fractionation	PTV dose values increased by values bigger than 100cGy for VMAT and IMRT plans
LINAC mechanical	Collimator angle errors	Collimator angle values increased by 10-15 degrees
General radiotherapy plan errors	Bolus usage	Bolus involvement to the plans that bolus was not prescribed and bolus absence to the plans involved the prescription of bolus

Eclipse (versions 11 and 15) treatment planning system database and amended with information from the electronic patient dossier (EPD). A description of all the variables used as well as with representative examples can be found in Table I.

B. Variable Mapping

The numerical variables of the dataset were mapped to the nearest value in the corresponding variable from Luk *et al.* For the categorical variables, such as anatomic tumor location, we mapped the values from our dataset to the matching values in the corresponding variable. If there was more than one matching value (e.g., the variable *T*_stage contains the values 1a, 1A, and T1a), we selected the one with the highest marginal probability in the model (i.e., the most common occurrence in the original training dataset).

C. Errors

Reports of errors and near-misses that happened in Maastro clinic (The Netherlands) related to radiotherapy were collected and validated from the prevention and recovery information system for monitoring and analysis (PRISMA) database [26]. After the assessment of the 19 054 treatment plans of the 5238 patients, we encountered five radiotherapy treatment plan errors reported that were checked manually. One of the errors was related to a wrong table angle, two errors were related to an incorrect planning tumor volume (PTV) dose and the remaining two errors were related to the usage of the bolus. Since our goal is to replace or support these manual and time-consuming checks with the introduction of BNs, we simulated errors in 3% of the plans following instructions of experts in the area.

These errors can be categorized into four main types: 1) patient positioning; 2) prescription level; 3) LINAC mechanical; and 4) general radiotherapy plan errors. The patient positioning error category consisted of the LINAC table rotation errors simulation errors with a values bigger than 10°. In the category of prescription level errors, differences between the prescribed dose to the PTV and the dose per fraction were evaluated. Specifically, we simulated errors with values bigger than 100 cGy planned dose to the PTV on VMAT and IMRT plans of 15 and 20 fractions. Errors regarding the LINAC collimator angle were simulated and included into the LINAC mechanical errors. In this category, the simulated errors collimator angle values were increased by 10°–15°. Under the category of the generic radiotherapy plan errors, we simulated errors for whether the usage of bolus or not was included. In Table II, you can find different categories and the description of the errors. The selection of the above-mentioned simulated errors was based on the reported and manually checked errors of the PRISMA database (table rotation, incorrect PTV dose, and bolus usage) and the suggestions of manually checked errors (collimator angle) from the radiotherapy technologists (RTTs) of Maastro Clinic.

D. Evaluation

We used the Java application programming interface (API) of Hugin Researcher 7.4 [27] to load the network provided by the authors and calculate the relevant probabilities. Following the instructions in the original article, for each case, we instantiated the variables Anatomic_tumor_loc, *T*_Stage, *M*_Stage, and *N*_Stage, and Treatment_Intent and calculated the probabilities of the rest of the variables. Each probability *P* was compared against a threshold *T* that designated whether

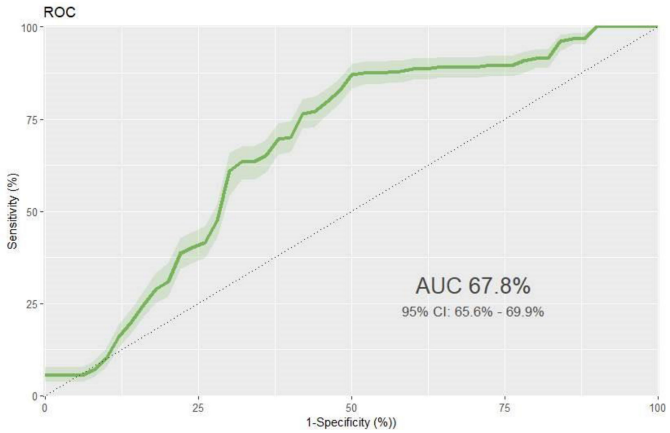

Fig. 1. ROC curve for the external validation dataset.

that parameter should be flagged as correct or as an error. Setup_Device variables were excluded, since these were not available in our database.

In order to compare the performance of the model reported by its authors with its performance in our dataset, we plotted the receiver-operating characteristic curve (ROC) and calculated the area under the curve (AUC), which provides an estimate of the discriminative power of the model. We plotted the ROC and calculated the AUC of the whole dataset (i.e., all variables combined) as well as for each of the variables where we simulated errors: collimator angle, table angle, gantry angle, PTV dose, and bolus. We used the ROC and calculated the AUC and its confidence intervals (CIs) using the *R* language (version 3.6.1) and the "classifierplots" package.

We also performed an analysis to assess the usefulness of the model in a clinic as an alert system that helps reduce the need for manual checks. As such, it would be only of added value if it could detect almost all errors (i.e., sensitivity $\geq$ 95%) with a reasonable positive predictive value (PPV, i.e., the probability that an instance flagged as an error is actually an error). Therefore, we undertook scenario analyses to calculate the model's PPV for different sensitivities and different prevalence of errors (since the PPV depends on how frequently errors occur in clinical practice and the prevalence is unknown). We did not assess calibration because the model's output is not meant to be interpreted as a probability.

The source code of our analysis is available at https://gitlab.com/UM-CDS/projects/ext-val-bn-rt-plan-qa.

III. Results

Fig. 1 shows the ROC curve for all the variables used in the external validation. The model achieved an AUC of 67.8% (95% CI, 65.6%–69.9%) when considering all variables together.

Table III shows the AUCs for the six types of simulated errors. The discriminative performance of the model is very high for the table rotation errors ("Table Angle" variable) achieving an AUC of 90.4% (95% CI, 87.1%–93.5%). For the category of the simulated errors related to the bolus, gantry angle, and the collimator angle, the model performs worse with AUCs of 75.6 % (95% CI, 71.3%–79.9%), 67% (61%–72.7%), and 69.6% (66.3%–73.1%), respectively. However, the BN fails to detect the errors comprising a difference between the prescribed dose to the PTV and the dose per fraction, resulting in an AUC of 54.5% (49.3%–59.4%).

TABLE III
AUCs for Different Types of Errors

Type of error	Mean	95% CI
Bolus	75.6	71.3 - 79.9
Collimator angle	69.6	66.3 - 73.1
Table angle	90.4	87.1 - 93.5
PTV dose	54.5	49.3 - 59.4
Gantry angle	67.0	61.0 - 72.7
Overall	**67.8**	**65.6 – 69.9**

TABLE IV
Percentage of Flagged Values and PPV for Different Combinations of Sensitivity and Prevalence of Errors

Sensitivity	Percentage of flagged values	PPV (%)		
		Prevalence of errors		
		0.1%	1%	3%
95%	68%	0.14	1.39	4.15
97%	78%	0.13	1.25	3.73
99%	79%	0.13	1.26	3.75

The results of our analysis regarding the usefulness of the model as an alert system are shown in Table IV, which includes the probability threshold at which different levels of high sensitivities are achieved and the resulting percentage of values flagged as errors and PPVs. According to our analyses, the model would flag as possible errors 84%, 89%, and 90% of the values in order to detect 95%, 97%, and 99% of errors, respectively. This implies that human technicians would still need to manually review almost all values to check whether they are correct. For these high sensitivity levels, the PPV, or the probability that a value flagged as an error is actually an error, was not significantly higher than the error prevalence itself.

Table V includes some of the cases from the external validation dataset where the model missed and detected errors. We selected the missed errors from those plans containing errors for which the model estimated a probability higher than the median probability in the test set for the variable that contained the error. Detected errors were selected from those plans containing errors for which the model estimated a probability lower than the 3rd percentile probability in the test set for the variable that contained the error. The analysis of patterns in the cases where the model succeeded and failed could potentially lead to insights to guide retraining and fine-tuning the process in the future.

TABLE V
SELECTION OF MISSED ERRORS (ESTIMATED PROBABILITY HIGHER THAN THE MEDIAN) AND DETECTED ERRORS (ESTIMATED PROBABILITY LOWER THAN THE 3RD PERCENTILE)

Missed errors					
Anatomic tumor location	**TNM Stage**			**Error**	**Erroneous value**
PROSTATE GLAND	T2 a	N0	M0	Bolus should be present	None
CERVIX	T1 b1	N0	M0	Gantry angle should start at 170	182
SKIN	T2	N0	M0	PTV dose should be 4000 cGy	3600
LUNG	T1c	N0	M0	Table angle should have been 10	0
HEAD /FACE/ NECK	T1	N0	M0	Radiation type should be 10X	6X
Detected errors					
Anatomic tumor location	**TNM Stage**			**Error**	**Erroneous value**
LUNG	T4	N3	M1 c	There should be no bolus	*custom
BREAST FEMALE	T2	N0	M0	Gantry angle should start at 168	179
BREAST FEMALE	T4 d	N1	M0	PTV dose should be 4500 cGy	4005
PROSTATE GLAND	NULL	NULL	M0	Table angle should have been 0	5
ABDOMEN	T3	N2	M1	Radiation type should be 6X	10X

IV. DISCUSSION

We have performed an external validation of a BN for error detection in radiotherapy plans described in [20] using data routinely collected at Maastro clinic. The results show that the model's performance is significantly deteriorated when using it outside of the environment it was developed in. We have also shown that the performance of the model varies heavily for different types of errors. We undertook an analysis that shows that in order to achieve a high sensitivity, the model needs to flag almost all values as potential errors, which reduces its usefulness as an alert system.

The deterioration in the performance of the model in our external validation might be caused by differences in radiotherapy practices between the two clinics and limitations in the implementation of the original model. For example, the institution from which the data to train the model originated uses Elekta's MOSAIQ oncology information system, while Maastro uses Varian's ARIA (Eclipse treatment planning system). On the other hand, the poor performance of the model detecting PTV dose errors could be caused by differences in institutional preferences on dose prescriptions and fractionation schedules. For example, in our institute, hypofractionation ($>$2 Gy per fraction) is frequently applied in prostate cancer patients, while in the original dataset used to train the model, a more conventional treatment schedule was used. The model flagging fractionation schedules different to those in the original institution as errors is likely to be a consequence of training a model in a single institution. However, it is arguable to which extent such models need to be generalizable (e.g., able to accept different fractionation schedules) and to which extent they should be adjusted to the implementing clinic (e.g., to deliberately flag as errors fractionation schedules different to the clinic's) through a commissioning process [28].

Another potential source of model performance deterioration is limitations in the model's development. For example, some categorical variables in the model contain redundant values (e.g., the variable *T*_Stage contains the values "1a," "1A," and "T1a") and numerical variables often contain a high number of values (e.g., more than 200 states). This in turn led to conditional probability tables (CPTs) with a high number of parameters, as the number of probabilities in a CPT grows exponentially with the number of states in each variable (e.g., the CPT for Number_of_Rxs contains more than 20 million probabilities). Since these parameters need to be estimated from data, the higher the number of parameters, the higher the number of samples required to learn these parameters. Options to alleviate the issue by reducing the number of values in each variable include removing redundant values, discretizing numeric variables, and grouping values that are similar or equivalent when considering the task at hand. In addition, the evaluation of the network as proposed by Luk *et al.* [21] considers the probability of each plan parameter independently, conditioned on the diagnostic variables and the treatment intent. This prevents the model from being able to detect erroneous combinations of plan parameters, such as a wrong value for Total_Fraction given a particular Dose_Per_Fraction.

It is worth assessing whether using a single probability threshold to determine whether to flag a value as an error or not is ideal. There is a high variance in the number of states or categories across different variables and probabilities

tend to be lower the higher the number of states. Therefore, adjusting the threshold per variable to reflect this could lead to improved performance.

There is also room for improvement in the handling of missing data. Many variables contain a special value to reflect missing data (e.g., "NULL"). This approach has been shown to lead to suboptimal results and is unnecessary in this case given that the algorithm used to learn the probabilities, the expected maximization (EM) algorithm, is especially suited to handle missing data [29]. Moreover, BNs are well capable of dealing with missing data when queried for probabilities (i.e., inference).

Moreover, the model was trained using data from a single institution. This is a common practice, but one that leads to models that often do not generalize well outside of the environment in which they were developed. Our results offer yet another example of the importance of using data from multiple sources (e.g., different clinics across the world) when training and testing models to achieve generalizability. This is not easy to achieve because ethical and legal barriers prevent sharing of privacy-sensitive data. However, the recently proposed federated learning paradigm [30] and related initiatives such as the Personal Health Train [31] aim to provide a framework where learning from multiple sites becomes straightforward. Another barrier to combining data from multiple institutions is the differences in the way different institutions encode the data. The findable, accessible, interoperable, and reusable (FAIR) data [32] principles establish a series of guidelines to make data interoperable, specifically by using publicly available ontologies for the creation of a semantic Web model. Such ontologies already exist for radiation oncology and radiotherapy [33]–[36].

Our external validation suffers from a number of limitations. The most important limitation is that while the information about the plans used in the validation is real, the errors are simulated. As explained in the methods section, after analyzing the database used to log misses and near-misses, we only found five errors related to radiotherapy planning. This is likely because technicians check every plan manually before and correct it before the plans are approved for treatment execution or because some errors go undetected. As a consequence, we were forced to simulate errors. Considering how much the model's performance varies across different types of errors, differences between the simulated and actual error distributions could lead to biased overall performance estimates. We mitigated this risk by simulating the errors partly based on the errors encountered in the database, and partly also by simulating the kind of errors that are manually corrected according to experienced technicians' feedback. Another limitation of our external validation is that our dataset was missing the information about the setup or immobilization devices (e.g., breast board and head rest) used during radiotherapy. As a consequence, we could not validate the performance of the model detecting errors in these variables. Finally, we did not assess the model's ability to detect errors that might have gone unnoticed in the clinic. In principle, by sacrificing sensitivity, one could use the model to try to flag a few errors that could otherwise go unnoticed with high specificity. However, this could be potentially dangerous because the existence of such a system could give a false sense of security to technicians unaware that by sacrificing sensitivity, most errors would go undetected.

The above-mentioned limitations in combination with the different radiotherapy treatment planning software between the two clinics (Mosaiq in Washington versus ARIA Eclipse in Maastro) and the LINAC models (Elekta in Washington versus Varian in Maastro) contributed to the relatively low performance of the model in the external validation. To further investigate the root cause of the low performance of the model in the validation cohort, we aim to address the limitations mentioned in the discussion and train the model in Maastro clinic as a next step of a future study.

The findings of our external validation suggest that the model is not yet ready to be useful in clinical practice in institutions different to its original. However, we believe that if the limitations identified in this external validation are successfully addressed, such a model could lead to a reduction in the cost of radiotherapy planning and increase its safety.

V. Conclusion

We have performed an external validation of a BN for error detection in radiotherapy plans proposed by Luk *et al.* [21], by testing the performance of the model in actual plans delivered in Maastro clinic with simulated errors. The results show that the performance of the model proposed by Luk *et al.* [21] significantly deteriorated when applied in an environment different from the source institution where it was developed (AUC of 65% versus 89%). The performance of the model varied widely for different types of errors (from 99.5% for table angle errors to 39.2% for PTV dose errors). This result shows the importance of external validations and the advantages of developing models using data from more than one institution. We analyzed the apparent limitations of the model (data preprocessing, handling of missing data, and model evaluation) and we have proposed actions to overcome them.

Acknowledgment

The authors wish to thank Nico Lustberg (Business Intelligence Developer), for helping with the extraction of the data from the ARIA oncology information system at Maastro.

References

[1] G. Delaney, S. Jacob, C. Featherstone, and M. Barton, "The role of radiotherapy in cancer treatment: Estimating optimal utilization from a review of evidence-based clinical guidelines," *Cancer*, vol. 104, no. 6, pp. 1129–1137, 2005, doi: 10.1002/cncr.21324.

[2] U. Ringborg *et al.*, "The Swedish council on technology assessment in health care (SBU) systematic overview of radiotherapy for cancer including a prospective survey of radiotherapy practice in Sweden 2001—Summary and conclusions," *Acta Oncologica*, vol. 42, nos. 5–6, pp. 357–365, 2003, doi: 10.1080/02841860310010826.

[3] A. C. Begg, F. A. Stewart, and C. Vens, "Strategies to improve radiotherapy with targeted drugs," *Nat. Rev. Cancer*, vol. 11, no. 4, pp. 239–253, 2011, doi: 10.1038/nrc3007.

[4] G. C. Barnett *et al.*, "Normal tissue reactions to radiotherapy: Towards tailoring treatment dose by genotype," *Nat. Rev. Cancer.*, vol. 9, no. 2, pp. 134–142, 2009, doi: 10.1038/nrc2587.

[5] T. Lustberg *et al.*, "Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer," *Radiotherapy Oncol.*, vol. 126, no. 2, pp. 312–317, 2018, doi: 10.1016/j.radonc.2017.11.012.

[6] N. Kim, J. S. Chang, Y. B. Kim, and J. S. Kim, "Atlas-based auto-segmentation for postoperative radiotherapy planning in endometrial and cervical cancers," *Radiat. Oncol.*, vol. 15, no. 1, p. 106, 2020, doi: 10.1186/s13014-020-01562-y.

[7] S. Cilla *et al.*, "Template-based automation of treatment planning in advanced radiotherapy: A comprehensive dosimetric and clinical evaluation," *Sci. Rep.*, vol. 10, no. 1, p. 423, 2020, doi: 10.1038/s41598-019-56966-y.

[8] *WHO Radiotherapy Risk Profile-Technical Manual*. [Online]. Available: https://www.who.int/patientsafety/activities/technical/radiotherapy_risk_profile.pdf

[9] B. Fraass, K. Doppke, M. Hunt, G. Kutcher, G. Starkschall, R. Stern, and J. V. Dyke, "American association of physicists in medicine radiation therapy committee task group 53: Quality assurance for clinical radiotherapy treatment planning," *Med. Phys.*, vol. 25, no. 10, pp. 1773–1829, 1998, doi: 10.1118/1.598373.

[10] D. Thwaites, P. Scalliet, J. W. Leer, and J. Overgaard, "Quality assurance in radiotherapy," *Radiotherapy Oncol.*, vol. 35, no. 1, pp. 61–73, 1995, doi: 10.1016/0167-81409501549-V.

[11] (Dec. 2013). *Radiotherapy Errors and Near Misses Data Report*. Accessed: Nov. 2015. [Online]. sAvailable: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/549847/radiotherapy_errors_and_near_misses_data_report.pdf

[12] (Sep. 2015). *Unintended Overexposure of a Patient During Radiotherapy Treatment at the Edinburgh Cancer Centre*, [Online]. Available: https://www.gov.scot/publications/unintended-overexposure-patient-during-radiotherapy-treatment-edinburgh-cancer-centre-september/

[13] C. J. A. Wolfs, R. A. M. Canters, and F. Verhaegen, "Identification of treatment error types for lung cancer patients using convolutional neural networks and EPID dosimetry," *Radiotherapy Oncol.*, vol. 153, pp. 243–249, Oct. 2020, doi: 10.1016/j.radonc.2020.09.048.

[14] L. Vandewinckele *et al.*, "Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance," *Radiotherapy Oncol.*, vol. 153, pp. 55–66, Dec. 2020, doi: 10.1016/j.radonc.2020.09.008.

[15] D. Yang and K. L. Moore, "Automated radiotherapy treatment plan integrity verification: Plan checking using PINNACLE scripts," *Med. Phys.*, vol. 39, no. 3, pp. 1542–1551, 2012, doi: 10.1118/1.3683646.

[16] B. Sun *et al.*, "Initial experience with TrueBeam trajectory log files for radiation therapy delivery verification," *Pract. Radiat. Oncol.*, vol. 3, no. 4, pp. e199–e208, 2013, doi: 10.1016/j.prro.2012.11.013.

[17] J. Xia, C. Mart, and J. Bayouth, "A computer aided treatment event recognition system in radiation therapy: Error detection in radiation therapy," *Med. Phys.*, vol. 41, no. 1, Art. no. 011713, 2013, doi: 10.1118/1.4852895.

[18] C. Holdsworth *et al.*, "Computerized system for safety verification of external beam radiation therapy planning," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 98, no. 3, pp. 691–698, 2017, doi: 10.1016/j.ijrobp.2017.03.001.

[19] T. Halabi and H. Lu, "Automating checks of plan check automation," *J. Appl. Clin. Med. Phys.*, vol. 15, no. 4, pp. 1–8, 2014, doi: 10.1120/jacmp.v15i4.4889.

[20] M. Hussein, B. J. M. Heijmen, D. Verellen, and A. Nisbet, "Automation in intensity modulated radiotherapy treatment planning—A review of recent innovations," *Brit. J. Radiol.*, vol. 91, no. 1092, 2018, Art. no. 20180270, doi: 10.1259/bjr.20180270.

[21] S. M. H. Luk *et al.*, "Characterization of a Bayesian network-based radiotherapy plan verification model," *Med. Phys.*, vol. 46, no. 5, pp. 2006–2014, 2019, doi: 10.1002/mp.13515.

[22] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Rev. 2*. San Mateo, CA, USA: Morgan Kaufmann, 2008.

[23] R. E. Neapolitan, *Learning Bayesian Networks*. Upper Saddle River, NJ, USA: Prentice-Hall, 2004.

[24] E. Kyrimi, S. McLachlan, K. Dube, M. R. Neves, A. Fahmi, and N. Fenton. (Feb. 28, 2020). *A Comprehensive Scoping Review of Bayesian Networks in Healthcare: Past, Present and Future*. Accessed: Nov. 5, 2020. [Online]. Available: http://arxiv.org/abs/2002.08627

[25] C. S. Collins *et al.*, "External validation of multivariable prediction models: A systematic review of methodological conduct and reporting," *Med. Res. Methodol.*, vol. 14, no. 1, p. 40, 2014, doi: 10.1186/1471-2288-14-40.

[26] W. Vuuren, V. T. W. Schaaf, and V. Der. *The Development of an Incident Analysis Tool for the Medical Field*. [Online]. Available: https://research.tue.nl/en/publications/the-development-of-an-incident-analysis-tool-for-the-medical-fiel

[27] S. K. Andersen, K. G. Olesen, F. V. Jensen, F. Jensen, Shafer and G. Pearl, Eds., "HUGIN—A shell for building belief universes for expert systems," in *Proc. Reading Uncertainty*, 1990, pp. 332–337.

[28] G. Mahadevaiah, P. Rv, I. Bermejo, D. Jaffray, A. Dekker, and L. Wee, "Artificial intelligence-based clinical decision support in modern medical physics: Selection, acceptance, commissioning, and quality assurance," *Med. Phys.*, vol. 47, no. 5, pp. e228–e235, 2020, doi: 10.1002/mp.13562.

[29] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc. B (Methodol.)*, vol. 39, no. 1, pp. 1–22, 1977, doi: 10.1111/j.2517-6161.1977.tb01600.x.

[30] M. J. Sheller *et al.*, "Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data," *Sci. Rep.*, vol. 10, no. 1, 2020, Art. no. 12598, doi: 10.1038/s41598-020-69250-1.

[31] O. Beyan *et al.*, "Distributed analytics on sensitive medical data: The personal health train," *Data Intell.*, vol. 2, nos. 1–2, pp. 96–107, 2020, doi: 10.1162/dint_a_00032.

[32] M. D. Wilkinson *et al.*, "The FAIR guiding principles for scientific data management and stewardship," *Sci. Data*, vol. 3, no. 1, 2016, Art. no. 160018, doi: 10.1038/sdata.2016.18.

[33] A. Traverso, J. van Soest, L. Wee, and A. Dekker, "The radiation oncology ontology (ROO): Publishing linked data in radiation oncology using semantic Web and ontology techniques," *Med. Phys.*, vol. 45, no. 10, pp. e854–e862, 2018, doi: 10.1002/mp.12879.

[34] *National Cancer Institute Thesaurus Ontology*. Accessed: Nov. 4, 2021. [Online]. Available: https://bioportal.bioontology.org/ontologies/NCIT

[35] *Radiation Oncology Ontology (ROO)*. Accessed: Nov. 4, 2021. [Online]. Available: https://bioportal.bioontology.org/ontologies/ROO

[36] M. H. Phillips *et al*. "Ontologies in radiation oncology," *Physica Medica*, vol. 72, pp. 103–113, Apr. 2020, doi: 10.1016/j.ejmp.2020.03.017.