Classification of various sources of error in range assessment using proton radiography and neural networks in head and neck cancer patients

Seller Oria, Carmen; Marmitt, Gabriel G; Both, Stefan; Langendijk, Johannes A; Knopf, Antje-Christin; Meijers, Arturs

Published in:
Physics in Medicine and Biology

DOI:
10.1088/1361-6560/abc09c

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

PAPER • OPEN ACCESS

Classification of various sources of error in range assessment using proton radiography and neural networks in head and neck cancer patients

To cite this article: Carmen Seller Oria *et al* 2020 *Phys. Med. Biol.* **65** 235009

View the article online for updates and enhancements.

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

OPEN ACCESS

RECEIVED
2 December 2019

REVISED
15 September 2020

ACCEPTED FOR PUBLICATION
13 October 2020

PUBLISHED
20 November 2020

PAPER

Classification of various sources of error in range assessment using proton radiography and neural networks in head and neck cancer patients

Carmen Seller Oria[1,*], Gabriel Guterres Marmitt[1], Stefan Both[1], Johannes A Langendijk[1], Antje C Knopf[1] and Arturs Meijers[1,2]

[1] Department of Radiation Oncology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
[2] Varian Medical Systems, Inc., Palo Alto, California, United States of America
[*] Author to whom any correspondence should be addressed

E-mail: c.seller.oria@umcg.nl

Keywords: range uncertainties, proton radiography, neural networks, range probing

Abstract

This study evaluates the suitability of convolutional neural networks (CNNs) to automatically process proton radiography (PR)-based images. CNNs are used to classify PR images impaired by several sources of error affecting the proton range, more precisely setup and calibration curve errors. PR simulations were performed in 40 head and neck cancer patients, at three different anatomical locations (fields A, B and C, centered for head and neck, neck and base of skull coverage). Field sizes were 26 × 26cm^2 for field A and 4.5 × 4.5cm^2 for fields B and C. Range shift maps were obtained by comparing an unperturbed reference PR against a PR where one or more sources of error affected the proton range. CT calibration curve errors in soft, bone and fat tissues and setup errors in the anterior–posterior and inferior–superior directions were simulated individually and in combination. A CNN was trained for each type of PR field, leading to three CNNs trained with a mixture of range shift maps arising from one or more sources of range error. To test the full/partial/wrong agreement between predicted and actual sources of range error in the range shift maps, exact, partial and wrong match percentages were computed for an independent test dataset containing range shift maps arising from isolated or combined errors, retrospectively. The CNN corresponding to field A showed superior capability to detect isolated and combined errors, with exact matches of 92% and 71% respectively. Field B showed exact matches of 80% and 54%, and field C resulted in exact matches of 77% and 41%. The suitability of CNNs to classify PR-based images containing different sources of error affecting the proton range was demonstrated. This procedure enables the detection of setup and calibration curve errors when they appear individually or in combination, providing valuable information for the interpretation of PR images.

1. Introduction

Proton therapy is a promising radiotherapy technique in terms of healthy tissue sparing. The deposited dose increases as protons slow down through matter, culminating in the Bragg peak, a steep increase and fall-off after which protons rapidly stop (Paganetti 2012). In comparison with conventional photon radiotherapy, proton therapy enables more advantageous dose conformity with lower integral dose, and more effective sparing of healthy tissues and organs at risk.

However, uncertainties in the proton range from various sources can lead to under or over-dosage of the target volume and/or undesired over-dosage in nearby healthy tissue. For this reason, proton therapy is still not used to its full potential (Paganetti 2012). One of the main research interests remains in the detection and mitigation of different uncertainties that are compromising the quality of proton treatments (Unkelbach *et al* 2007, Mumot *et al* 2010, Fredriksson *et al* 2011, Knopf and Lomax 2013, Parodi and Polf 2018).

The proton range (Newhauser and Zhang 2015) can be affected by anatomical changes, organ motion, and setup errors that lead to density changes in the beam path, which translate into variations of the delivered dose distribution. Furthermore the calibration curve for the conversion of CT numbers into density values, and consequently stopping power ratios, can be a source of range uncertainties (Paganetti 2012). Finally, changes in tissue composition can affect the range accuracy (Paganetti 2012). In this work, the feasibility to detect effects of calibration curve errors and setup errors by means of proton radiography (PR) is investigated making use of convolutional neural networks (CNNs).

Mumot *et al* suggested that PR, obtained by using a distribution of range probes, could be used for *in vivo* range verifications with few probes (Mumot *et al* 2010). In previous studies, PR was used with the aim to identify individual sources of error (Farace *et al* 2016a, 2016b). In addition, some studies investigated experimentally different body sites to perform range probing for positioning purposes with a head phantom (Hammi *et al* 2018a, 2018b). Furthermore, the effects of setup errors in range assessment have been evaluated, concluding that they could be distinguished from other sources of error affecting the proton range (Deffet *et al* 2017). CT calibration curve errors have also been an object of investigation, pointing out that range probing (Mumot *et al* 2010) and PR can be used as efficient quality assurance tools, obtaining a direct measurement of the stopping power of tissues (Schneider and Pedroni 1994, Schneider *et al* 2005, Knopf and Lomax 2013, Doolan *et al* 2015). However, research to date has not resulted in a comprehensive PR tool that allows for automated and fast interpretation of PR images, identifying different sources of error affecting the proton range.

The broad applicability of artificial intelligence and its development is becoming increasingly present in medical imaging and radiation oncology research. Deep learning and CNNs have been employed for tasks such as organ recognition and segmentation, image processing and reconstruction, synthetic CT generation or prediction of dose distributions (Cha *et al* 2016, Jin *et al* 2017, Nguyen *et al* 2019, Thummerer *et al* 2020). To this end, neural networks are trained and tested with datasets acquired from several patients, containing images or image patches of diverse modalities such as magnetic resonance, ultrasound or CT scans, according to the task that they are meant to perform. There is as well a wide variety of architectures, ranging from conventionally defined networks like U-net to more customized architectures (Sahiner *et al* 2019).

In this study we present a proof of concept which aims to evaluate the suitability of CNNs to classify PR based images impaired by different sources of error affecting the proton range. To our knowledge, this is the first time that the use of CNNs has been explored for the interpretation of PR images affected by various sources of error.

2. Materials and methods

Range shift maps obtained from PR simulations in 40 head and neck cancer patients were interpreted by means of CNNs. Three PR fields of different size and at different anatomical regions were simulated. A dedicated CNN was employed to classify range shift maps arising from each type of PR field.

2.1. PR simulations

PR simulations were performed in openREGGUI (openreggui.org), a Matlab open source toolbox (Farace *et al* 2016a, Deffet *et al* 2017). It enables the simulation of PRs as those acquired by a Giraffe multi-layer ionization chamber (MLIC) (IBA Dosimetry, Schwarzenbruck, DE). The algorithm to produce PR simulations takes as input an MLIC measurement in air, and scales the integral depth dose curves (IDDs) in the beam axis according to the water equivalent thickness (WET) of the patient, as calculated from the provided CT image set. The algorithm accurately accounts for range mixing by convoluting a Gaussian kernel with the IDDs in air, given the beam sigma = 3.5 mm (Farace *et al* 2016a, Deffet *et al* 2017). Such simulations represent realistic measurement noise, given that the reference IDD is an actual MLIC measurement.

PR simulations were performed with pencil beams of 210 MeV, from a gantry angle of 270 degrees and with a pencil beam spacing of 1 mm. Planning CT scans from 40 different head and neck cancer patients previously treated in our facility were included in this study.

2.2. PR fields

PR fields were simulated to evaluate the performance of the method in three different anatomical regions and for different PR field sizes. A large PR field of 26×26 cm^2 was selected to cover entirely the head and neck region of the patient. Additionally, two small PR fields of 4.5×4.5 cm^2 centered in the neck and in the base of the skull were simulated. The three PR fields are respectively referred to as ‘field A’, ‘field B’ and ‘field C’. Figure 1 illustrates their location in an example patient geometry.

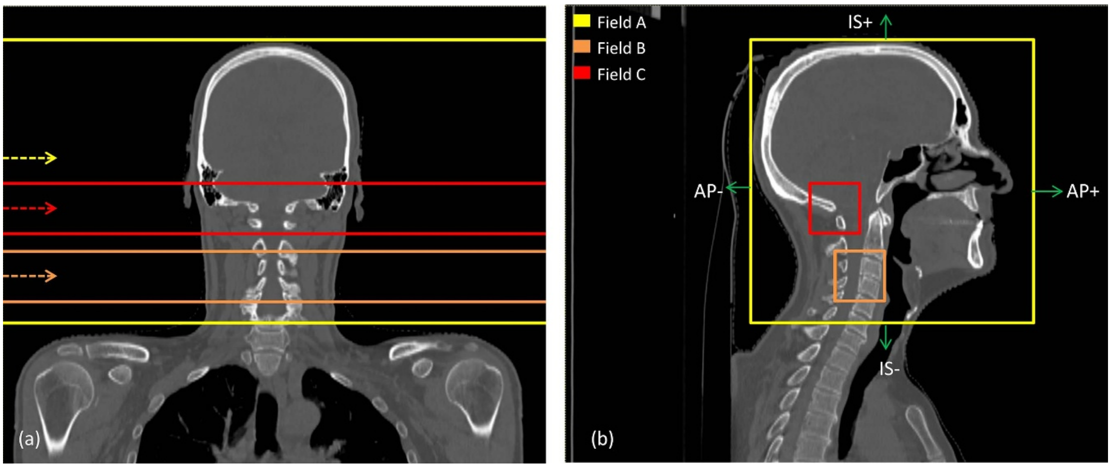

Figure 1. Coronal (a) and sagittal (b) views of an example patient geometry. The edges of PR fields A, B and C are highlighted in yellow, orange and red respectively. In the coronal view, dashed arrows indicate the beam direction (from left to right), meaning that the MLIC is located at the right side of the CT. In the sagittal view the beams enter towards the patient, thus the MLIC is behind the CT. Positive and negative setup errors in anterior-posterior and inferior-superior directions (AP+, AP-, IS+, IS-) are depicted by green arrows in the sagittal view for field A. The length of the arrows is illustrative, and it does not express the magnitude of the setup error.

Table 1. Generic unperturbed calibration curve expressed in terms of Hounsfield units (HUs) and density (ρ) in g cm^{-3}. Specific density values were subject to perturbations, leading to perturbed calibration curves. Perturbations in the fat, soft and bone tissue correspond to alterations in density values colored in blue, red and green respectively.

HU	−992	−976	−480	−96	48	128	528	976	1488	1824	2224	2640	2832	2833	3096
ρ [g cm^{-3}]	0.00121	0.00121	0.5	0.95	1.05	1.1	1.35	1.6	1.85	2.1	2.4	2.7	2.83	7.87	7.87

The size of field A was selected to guarantee head and neck coverage. The size of fields B and C was selected in accordance with other studies performed with an MLIC, to fully exploit the readout area of the MLIC detector (Farace *et al* 2016b).

The location of the three fields was chosen to cover a wide variety of tissues in the beam direction. Furthermore, the existence of lateral heterogeneities was favored, since it gives rise to different Bragg curves, which benefits setup error detection (Farace *et al* 2016b).

2.3. Error scenarios

A reference PR was simulated for each patient (CT scan), representing an unperturbed range measurement. Reference radiographies were compared against radiographies in which one or more sources of error affected the proton range. Calibration curve errors and setup errors were the sources of error under consideration.

2.3.1. Calibration curve errors

Table 1 shows the Hounsfield unit (HU) and density values of a generic calibration curve representing an unperturbed reference scenario. Calibration curve errors were simulated by perturbing density values of the calibration curve corresponding to the fat (F), soft (S) and bone (B) tissue regions (Ainsleya and Yeager 2014). Table 1 shows which specific density values were modified to simulate a perturbation in each type of tissue (F, S and B). Positive or negative signs were given to over or underestimations of density values, resulting in 6 calibration curve error categories (S+, B+, F+, S-, B- and F-).

The CNNs were trained with range shift maps arising from calibration curve errors (±3 to ± 7% in steps of 1% for S and F, ± 7 to ± 15% in steps of 2% for B). This choice is based on reported inter institutional variations of calibration curves (Peters *et al* 2019), and it has the purpose to mimic a clinical scenario, in which the calibration curve would show systematic errors. With this criteria, five calibration curves were produced for each calibration curve error category, giving rise to a total of 30 different calibration curves.

The dataset to test the CNNs performance contains random calibration curve errors within the same ranges of values (±3 to ± 7% in S and F, ± 7 to ± 15% in B), to evaluate whether the CNNs can detect calibration curve errors that were not shown during the training process.

2.3.2. Setup errors

In accordance with the setup error margin for head and neck patients adopted in our clinic (Van Dijk *et al* 2016), setup errors of 1, 2, 3 or 4 mm in the AP and IS directions were simulated by shifting the PR fields in

Table 2. Schematic of the error type combinations under consideration. Each setup error (AP+, AP-, IS+, IS-) can be combined with any type of calibration curve error (S+, S-, F+, F-, B+, B-).

	S+	S-	F+	F-	B+	B-
AP+	S+ AP+	S-AP+	F+ AP+	F-AP+	B+ AP+	B-AP+
AP-	S+ AP-	S-AP-	F+ AP-	F-AP-	B+ AP-	B-AP-
IS+	S+ IS+	S-IS+	F+ IS+	F-IS+	B+ IS+	B-IS+
IS-	S+ IS-	S-IS-	F+ IS-	F-IS-	B+ IS-	B-IS-

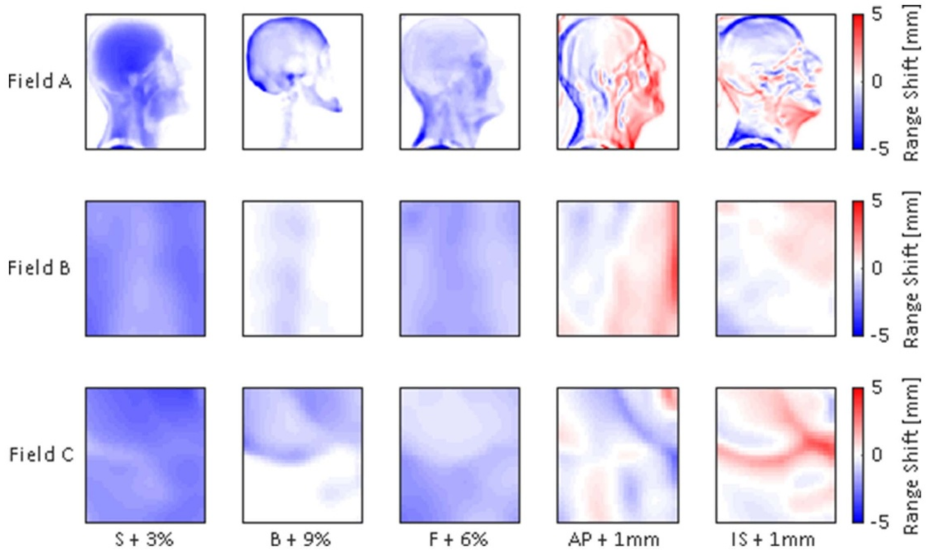

Figure 2. Five examples of range shift maps obtained for fields A, B and C. The maps correspond to isolated errors of different magnitudes. Out of the ten error categories, five are displayed (positive cases S+, B+, F+, AP+ and IS+). Range shifts are expressed in a scale from −5 mm to+5 mm.

the positive or negative direction (AP+, AP-, IS+, IS-), as depicted by the green arrows in figure 1. Both training and testing of the CNN were performed with setup errors from 1 to 4 mm in steps of 1 mm.

2.3.3. Isolated and combined errors

In total, ten different error categories were considered: S+, S-, B+, B-, F+, F-, AP+, AP-, IS+, IS-.

Besides looking at the influence of different sources of error separately, range shift maps arising from simulations with a combination of different types of error were analyzed (Chen *et al* 2012, Van Dijk *et al* 2016). Two errors at a time were simulated, and combinations of a CT calibration curve error with a setup error were considered. With such settings (six types of calibration curve errors combined with four types of setup errors), 24 different types of combinations are possible (table 2).

The comparison between reference PRs and PRs in which one or more sources of error affected the proton range was performed by computing the range shift of each IDD in the PR field. Range shifts were determined using the least squares method (minimizing the sum of squared differences between two IDDs (Farace *et al* 2016b)), and presented in the form of two dimensional maps, referred to as range shift maps, as shown in figure 2.

Figure 2 shows examples of range shift maps arising from five different isolated errors in fields A, B, and C. Positive and negative range shifts (red and blue pixels) represent IDDs with a higher or lower range in the perturbed PR with respect to the reference PR.

S+, B+ and F+ represent overestimations of density values, meaning that the WET crossed by the pencil beams was greater in the perturbed PR. The same line of thought applies to S-, B- and F-, which lead to the same patterns but positive range shifts.

Setup errors result in range shift maps composed of positive and negative range shifts (figure 2), whose arrangement depends on the direction of the setup error as displayed in figure 1.

The features observed in range shift maps corresponding to fields B and C in figure 2 can be better interpreted by observing the anatomical location of these fields within the patient (figure 1). For instance, the maps referred as ‘B+ 9%’ in fields B and C show clearly the location of vertebrae and base of skull within the maps.

Table 3. Features of the three CNNs used for fields A, B and C. The reported parameters are: number of convolutional layers with their corresponding number of filters and their size, number of fully connected layers (F.C) and their size, learning rate and batch size. The number of range shift maps in the training and validation set, as well as testing sets for isolated errors (TI) and combinations (TC) are displayed.

	Convol. layers	Number of filters	Filter size	F.C	F.C size	Learning rate	Batch size	Train+ val.	TI	TC
Field A	4	32,32,64,64	3	2	260	0.007	112	1120	200	240
Field B	1	32	2	2	1000	0.03	112	1120	200	240
Field C	1	32	2	2	1000	0.03	112	1120	200	240

2.4. Data analysis: the CNNs

The CNNs, implemented with TensorFlow (Abadi *et al* 2016), were used to recognize patterns in the range shift maps and classify them according to the different types of error that they contain.

2.4.1. CNN features

Every range shift map was labeled with a binary vector of ten elements corresponding to the ten different error categories. Each element or true value represents the presence (labeled as 1) or absence (labeled as 0) of each type of error in that range shift map. As mentioned in the previous section, a maximum of two types of error may be present in one range shift map. In machine learning, this is known as a multi-label classification problem.

Table 3 shows the features of the CNNs used to classify range shift maps from fields A, B, and C. The CNNs for fields B and C were assigned identical architectures.

Different number of convolutional layers with varying sizes were assigned to CNNs corresponding to fields A, B and C. The convolutional layers are followed by a max pooling layer and a rectifier linear unit (ReLu) activation function. The outcome of the pooling layer is flattened and introduced into two fully connected layers. The first fully connected layer uses a ReLu activation function. The output of the second fully connected layer is directed to a sigmoid activation function, which enables multi-label classification. Therefore, the prediction vector expresses the probability of each type of error independently in a range from 0 to 1 (Nielsen 2015, Chu *et al* 2017).

The cost function used during the training of the CNN is the mean sigmoid cross entropy between true values and predicted probabilities (Sokolova and Lapalme 2009, Liao *et al* 2016). The optimizer used is the gradient descendent algorithm. The training process stopped when the derivative of the cost function turned positive (right before the network starts over-fitting).

2.4.2. Training, validation and test datasets

CT scans of 30 patients were used to obtain range shift maps arising from one single type of error. Twenty PR simulations (two simulations of each error category with different magnitudes) were performed for each CT scan (patient). For simulations of combinations of errors, 40 patients were included. For each patient, 24 range shift maps were produced, corresponding with the number of possible error combinations as shown in table 2. In this way, each error category is equally represented.

The set of maps meant to train each CNN contained 1120 maps arising from 30 different patients: 400 maps from isolated errors and 720 maps from combinations of errors. For guidance purposes during the training process, 10% of those maps were randomly separated into a validation set.

Range shift maps from the remaining ten patients were assigned to the testing sets. In this way, none of the patients used for testing was previously introduced to the CNNs during the training process. As shown in table 3, one testing set contained 200 range shift maps arising from one error type (referred to as **test isolated or TI**), and another testing set had 240 range shift maps arising from combinations of two types of error (called **test combined or TC**).

2.4.3. Evaluation metrics

Once the training process was finished, a prediction vector was obtained for each range shift map in the testing sets. Each value of the prediction vector was split into two categories based on a threshold of 0.5. Predictions above/below the threshold were marked as a positive/negative prediction for that error. Thus, after applying the threshold, the resulting binary prediction vectors are ready to be compared against their corresponding true values (label).

The quality of the predictions was assessed by means of the exact, wrong and partial match percentages. An exact match occurs when all elements in a binarized prediction vector are identical to its corresponding true values. A wrong match occurs when none of the existing error types contained in a range shift map are detected. A partial match occurs when some of the elements in the prediction vector coincide with those in the corresponding label, but the binarized prediction vector and true value vector are not identical. Partial

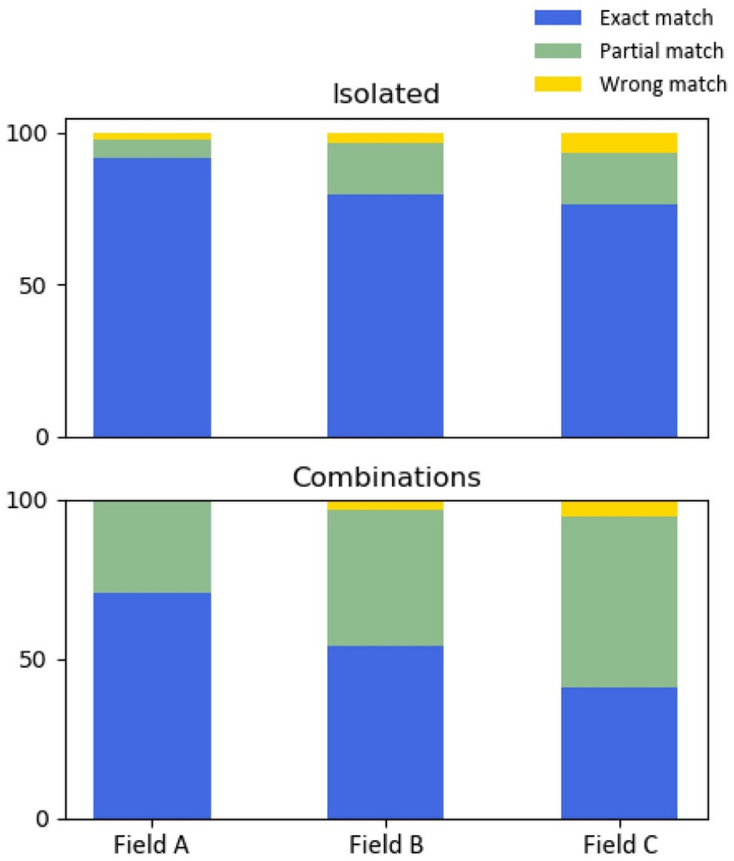

Figure 3. Exact, partial and wrong match percentages calculated for TI and TC (test datasets containing range shift maps arising from individual or combined errors).

matches can also occur in TI, if the CNNs predict two types of error for a range shift map that only contains one error type.

Exact, wrong and partial match percentages were computed both in TI and in TC, for the CNNs assigned to fields A, B and C. For TI, this evaluation was split by error categories (S±, B±, F±, AP±, IS±). For TC, exact, wrong and partial matches were computed for six groups of error combinations: S± with AP±, S± with IS±, B± with AP±, B± with IS±, F± with AP±, and F± with IS± .

3. Results and analysis

Figure 3 displays the performance evaluation for TI and TC, on fields A, B and C. The performance of the CNNs is superior when it classifies range shift maps with individual errors than when it classifies combinations. For isolated errors, exact match ratios for fields A, B and C are 92%, 80% and 78% respectively. For combinations of errors, exact match ratios are 71%, 54% and 41%. A small fraction of range shift maps were partially classified in TI (6%, 17% and 17% for fields A, B and C), while in TC the percentage of partial matches was larger (29%, 42% and 54%). A small number of the range shift maps was wrongly classified both in TI (2%, 3%, 5%) and in TC (0%, 4%, 6%).

A subsequent performance evaluation per error type is reported in figure 4 for TI. In the case of calibration curve errors, exact match percentages are nearly 100% in almost all cases. A slightly inferior outcome is obtained for errors of type F, where wrong matches of 10% were found in fields A and C. Setup error predictions show a higher rate of partial matches with respect to calibration curve errors. The most relevant cases appear in AP errors in field B (62% of partial matches) and in IS errors of field C (57% of partial matches). Wrong match percentages are seen for IS errors in field B and for AP errors in field C (17% and 22% respectively).

Figure 5 shows the outcome of the CNNs performance per types of combination (using TC). Field A shows superior performance with respect to fields B and C, since the exact match percentages lay in a range from 59% to 78% among all types of combination. Field B shows large performance variations across different types of error combinations. For instance, a combination of S with AP results in an exact/partial match percentage of 79/21% while F with AP lead to a ratio of 48/52%. For field C, the performance of the

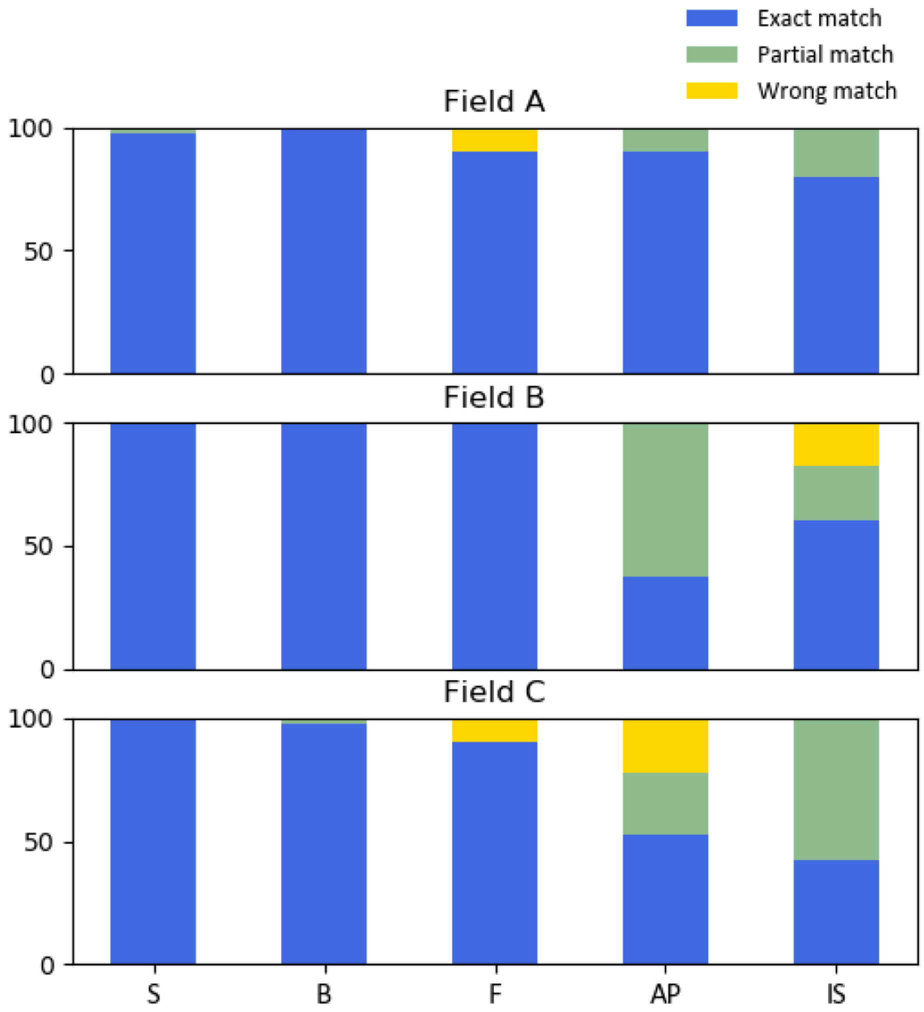

Figure 4. Exact, partial and wrong match percentages per class types for TI in fields A, B and C. S stands for S+ and S-, the same applies to the rest of categories.

CNN is similar for different types of combination, resulting in exact match percentages in a range from 32% to 49%.

4. Discussion

This study investigated the potential of CNNs to classify range shift maps arising from PRs in head and neck cancer patients. Calibration curve errors as well as setup errors of different magnitudes were simulated individually and in combination for 40 patients. Furthermore, three different locations were selected to perform PRs (fields A, B and C), and each of them was assigned to a dedicated CNN.

As shown in figure 3, CNNs in the three anatomical locations showed high performance (exact match percentages above 78%) when they process range shift maps containing one isolated type of error. This implies that the developed tool can indicate not only what is the origin of the existing error (setup error or calibration curve error), but also point out details about the direction of a setup error or the region of the calibration curve that is being affected, as well as whether densities are being over or under estimated.

Furthermore, the three CNNs were capable of detecting combinations of a calibration curve with a setup error contained in the same range shift map. In this case, the best outcome was found for field A, followed by fields B and C. Figure 3 shows that the PR field size plays an important role in the prediction of combined sources of error, since the exact match percentage in field A, which has the largest coverage of the patient anatomy (see figures 1 and 4), is at least 17% higher than in fields B and C.

The location of the PR fields is as well linked to the ability of the corresponding CNN to identify combinations of errors. For two fields of the same size (fields B and C), 13% better outcomes in terms of exact match percentages are obtained in field B, located at the neck. An optimal field location should be such that each type of error gives rise to a unique and characteristic pattern. Furthermore, the chosen anatomical region should contain the different tissues for which one desires to detect errors in the calibration curve.

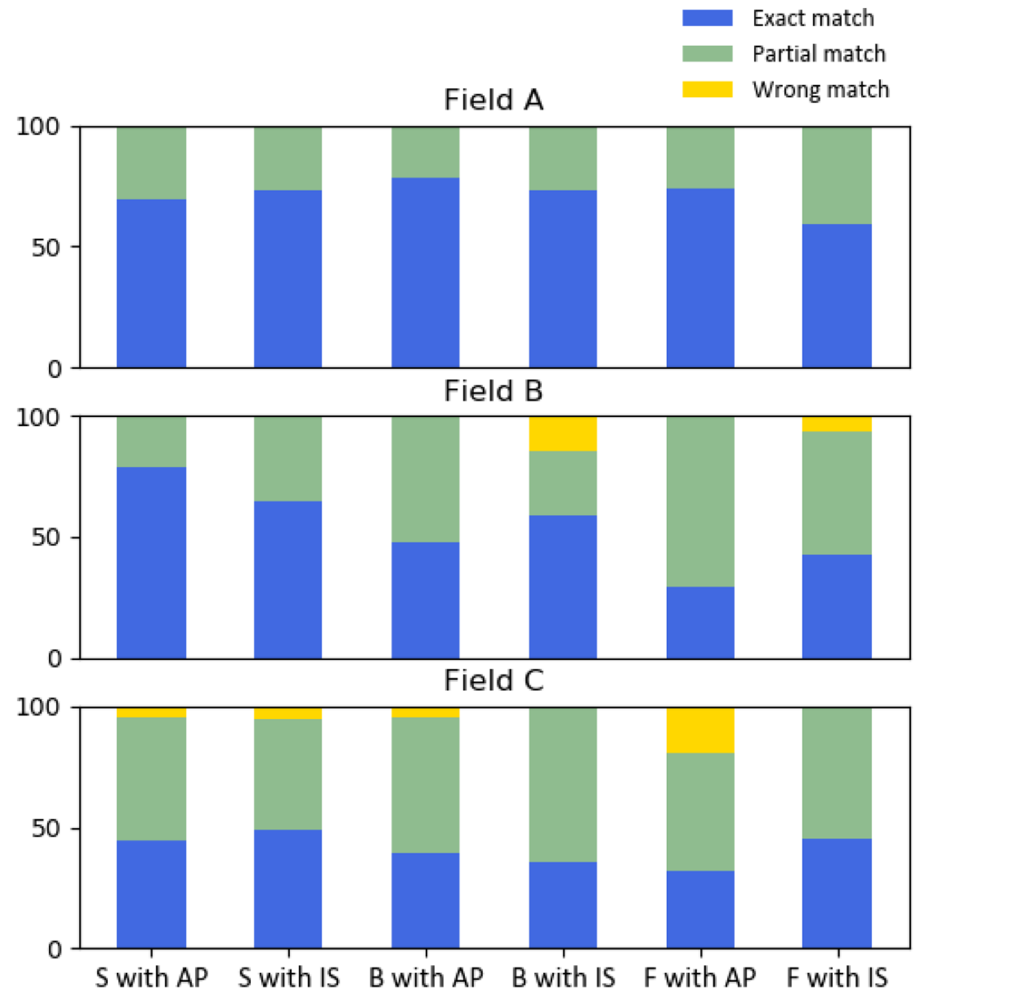

Figure 5. Exact, partial and wrong match percentages computed for separate groups of combinations in TC, for fields A, B and C. ‘ S with AP’ stands for S+ or S- with AP+ or AP-, the same applies to the rest of categories.

These requirements are not always fully achievable. For instance, the amount of fat tissue in the head and neck region is reduced in comparison with other tissues, leading to small range shifts. Furthermore, errors of type F result in range shift maps similar to those arising from errors of type S (figure 2), which imposes a challenge for the CNNs, especially when they are combined with a setup error.

A more detailed analysis was carried out in figures 4 and 5, where the performance of the CNNs was evaluated separately, per error type or per combination type. Figure 4 shows that calibration curve errors are easily detected when they appear individually. Setup errors are identified as well, although higher rates of partial matches are found. In those cases, the CNN predicted the existing setup error plus another type of error. In general, setup errors lead to larger range shifts than calibration curve errors (this is especially true for B and F types). The CNNs are acquainted to detect two error types in maps where setup errors have a dominant effect. This outcome does not impose a major limitation, since a PR could be acquired, the patient could be repositioned if dominant setup errors are detected by the CNN, and a second PR could be obtained to verify if there are remaining calibration curve errors. Local residual setup errors would always remain, so a compromise on the PR field size should be achieved.

In addition, CNNs could be complemented by other setup error mitigation algorithms such as the one developed by Deffet *et al* (2017), to help CNNs to achieve a more reliable detection of other sources of error affecting the proton range.

As previously mentioned, field size and location have a direct impact on the level of similarity between range shift maps of different categories. Wrong matches are more frequent for fields B and C (see figure 5), which cover a smaller anatomical region in the patients with respect to field A. Fields B and C are more likely to lead to similar range shift maps that belong to different categories. Furthermore, anatomical differences across patients can lead to unexpected features in range shift maps extracted from a small PR field.

Given that superior CNN performance was found for field A (26 × 26 cm^2) with respect to the rest of fields (4.5 × 4.5 cm^2), the field size and location should be further investigated and optimized. Furthermore, the developed tool could be extended to detect other types of isolated or combined errors, for example diagonal or rotational setup errors. In other anatomical regions like the thorax, the proton ranges are additionally influenced by organ motion due to respiration, which could be another source of error affecting the proton range to be explored.

CNNs are a common tool employed for image classification problems (Lu and Weng 2007). The CNN architectures reported in this study (table 3) are the outcome of an exhaustive parameter search, in which different number of convolutional layers, number and size of filters, fully connected layers, learning rates and batch sizes were explored. Future investigations could establish a comparison between different algorithms, evaluating if other image classification tools can increase the exact match percentages, especially when range shift maps arise from error combinations. For instance, hybrid solutions could be considered (Madjarov G and Gjorgjevikj D 2012).

Previous studies restricted the application of PR for range verification purposes to one source of error at a time, suggesting its applicability for patient alignment (2016a, Farace *et al* 2016b, 2018a, Hammi *et al* 2018b) or corrections in the CT calibration curve (Schneider and Pedroni 1994, Schneider *et al* 2005, Doolan *et al* 2015). This work presents a tool to detect several types of errors using a single PR acquisition. The proposed methodology could be extended to identify anatomical changes (for instance, due to weight variations) or breathing motion.

After an experimental validation of the proposed PR tool is carried out, the CNNs could be plugged into automated workflows to extract data from images with none or limited need of human interventions, and faster interpretation of quality control measurements on a patient specific basis could be performed. A PR of the patient in treatment position could be acquired right before the treatment fraction delivery, and range errors between the PR measurement and a PR simulated with the most recent CT of the patient would be computed. The resulting range shift map would be fed into a CNN that automatically analyzes the sources of eventual error affecting the proton range. The outcome of the CNN, combined with updated patient anatomy information, could assist on making decisions upon treatment quality and the need for plan adaptations. This study presents a further step towards future automated applications of PR for *in vivo* range verification.

5. Conclusion

The feasibility to classify PR images impaired by different types of errors affecting the proton range by means of CNNs was demonstrated. PR simulations were performed in head and neck cancer patients, introducing setup and calibration curve errors individually and in combination. The investigated methodology provides means of automated PR interpretation in the proton treatment quality control process.

Acknowledgments

This study was financially supported by a grant from the Dutch Cancer Society (KWF research project 11518), called ‘INCONTROL- Clinical Control Infrastructure for Proton Therapy Treatments’.

ORCID iDs

Carmen Seller Oria https://orcid.org/0000-0002-0785-2009
Gabriel Guterres Marmitt https://orcid.org/0000-0002-8486-7001

References

Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S and Irving G 2016 TensorFlow: A system for large-scale machine learning *12th USENIX Symp. Oper. Syst. Des. Implement. (OSDI) '16* pp 265–83

Ainsleya C G and Yeager C M 2014 Practical considerations in the calibration of CT scanners for proton therapy *J. Appl. Clin. Med. Phys.* **15** 202–20

Cha K H, Hadjiiski L, Samala R K, Chan H P, Caoili E M and Cohan R H 2016 Urinary bladder segmentation in CT urography using deep-learning convolutional neural network and level sets *Med. Phys.* **43** 1882–6

Chen W, Unkelbach J, Trofimov A, Madden T, Kooy H, Bortfeld T and Craft D 2012 Including robustness in multi-criteria optimization for intensity-modulated proton therapy *Phys. Med. Biol.* **57** 591–608

Chu W S, De La Torre F and Cohn J F 2017 Learning spatial and temporal cues for multi-label facial action unit detection *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognition, FG 2017 1st Int. Work. Adapt. Shot Learn. Gesture Underst. Prod. ASL4GUP 2017, Biometrics Wild, Bwild 2017, Heteroge* pp 25–32

Deffet S, Macq B, Righetto R, Vander Stappen F and Farace P 2017 Registration of pencil beam proton radiography data with X-ray CT *Med. Phys.* **44** 5393–401

Doolan P J, Testa M, Sharp G, Bentefour E H, Royle G and Lu H M 2015 Patient-specific stopping power calibration for proton therapy planning based on single-detector proton radiography *Phys. Med. Biol.* **60** 1901–17

Farace P, Righetto R, Deffet S, Meijers A and Vander Stappen F 2016a Technical note: A direct ray-tracing method to compute integral depth dose in pencil beam proton radiography with a multilayer ionization chamber *Med. Phys.* **43** 6405–12

Farace P, Righetto R and Meijers A 2016b Pencil beam proton radiography using a multilayer ionization chamber *Phys. Med. Biol.* **61** 4078–87

Fredriksson A, Forsgren A and Hårdemark B 2011 Minimax optimization for handling range and setup uncertainties in proton therapy *Med. Phys.* **38** 1672–84

Hammi A, König S, Weber D C, Poppe B and Lomax A J 2018a Patient positioning verification for proton therapy using proton radiography *Phys. Med. Biol.* **63** 245009

Hammi A, Placidi L, Weber D C and Lomax A J 2018b Positioning of head and neck patients for proton therapy using proton range probes: a proof of concept study *Phys. Med. Biol.* **63** 015025

Jin K H, Mccann M T, Froustey E and Unser M 2017 Deep convolutional neural network for inverse problems in imaging *IEEE Trans. Image Process.* **26** 4509–22

Knopf A C and Lomax A 2013 *In vivo* proton range verification: A review *Phys. Med. Biol.* **58** 131–60

Liao H, Li Y and Luo J 2016 Skin disease classification versus skin lesion characterization: achieving robust diagnosis using multi-label deep neural networks *Proc. Int. Conf. Pattern Recognit.* **0** 355–60

Lu D and Weng Q 2007 A survey of image classification methods and techniques for improving classification performance *Int. J. Remote Sens.* **28** 823–70

Madjarov G and Gjorgjevikj D 2012 Hybrid decision tree architecture utilizing local SVMs for multi-label classification *Int. Conf. on Hybrid Artif. Intell. Sys.* pp 1–12

Mumot M, Algranati C, Hartmann M, Schippers J M, Hug E and Lomax A J 2010 Proton range verification using a range probe: definition of concept and initial analysis *Phys. Med. Biol.* **55** 4771–82

Newhauser W D and Zhang R 2015 The physics of proton therapy *Phys. Med. Biol.* **60** R155–209

Nguyen D, Long T, Jia X, Lu W, Gu X, Iqbal Z and Jiang S 2019 A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning *Sci. Rep.* **9** 1–10

Nielsen M A 2015 *Neural Networks and Deep Learning* (San Francisco, CA: Determination Press)

Paganetti H 2012 Range uncertainties in proton therapy and the role of Monte Carlo simulations *Phys. Med. Biol.* **57** R99

Parodi K and Polf J C 2018 *In vivo* range verification in particle therapy *Med. Phys.* **45** e1036–50

Peters N *et al* 2019 Experimental assessment of inter-centre variation and accuracy in stopping power ratio prediction within the European Particle Therapy Network *ESTRO* **38** S348–9

Sahiner B, Pezeshk A, Hadjiiski L M, Wang X, Drukker K, Cha K H, Summers R M and Giger M L 2019 Deep learning in medical imaging and radiation therapy *Med. Phys.* **46** e1–36

Schneider U and Pedroni E 1994 Proton radiography as a tool for quality control in proton therapy *Med. Phys.* **22** 353–63

Schneider U, Pemler P, Besserer J, Pedroni E, Lomax A and Kaser-Hotz B 2005 Patient specific optimization of the relation between CT-Hounsfield units and proton stopping power with proton radiography *Med. Phys.* **32** 195–9

Sokolova M and Lapalme G 2009 A systematic analysis of performance measures for classification tasks *Inf. Process. Manag.* **45** 427–37

Thummerer A, Zaffino P, Meijers A, Marmitt G G, Seco J, Steenbakkers R J H M, Langendijk J A, Both S, Spadea M F and Knopf A-C 2020 Comparison of CBCT based synthetic CT methods suitable for proton dose calculations in adaptive proton therapy *Phys. Med. Biol.* **65** 095002

Unkelbach J, Chan T C Y and Bortfeld T 2007 Accounting for range uncertainties in the optimization of intensity modulated proton therapy *Phys. Med. Biol.* **52** 2755–73

Van Dijk L V, Steenbakkers R J H M, Ten Haken B, Van Der Laan H P, Van't Veld A A, Langendijk J A and Korevaar E W 2016 Robust Intensity Modulated Proton Therapy (IMPT) increases estimated clinical benefit in head and neck cancer patients *PloS One* **11** 1–15