Machine learning-based patient specific prompt-gamma dose monitoring in proton therapy

2013 Phys. Med. Biol. 58 4563

(http://iopscience.iop.org/0031-9155/58/13/4563)

OPEN ACCESS
IOP PUBLISHING PHYSICS IN MEDICINE AND BIOLOGY
Phys. Med. Biol. **58** (2013) 4563–4577 doi:10.1088/0031-9155/58/13/4563

Machine learning-based patient specific prompt-gamma dose monitoring in proton therapy

P Gueth[1,3], D Dauvergne[2], N Freud[1,3], J M Létang[1,3], C Ray[2], E Testa[2] and D Sarrut[1,3]

[1] Université de Lyon, CREATIS, CNRS UMR5220, Inserm U1044, INSA, F-69622 Lyon, France
[2] Université de Lyon, IPNL, CNRS/IN2P3 UMR5822, F-69622 Lyon, France
[3] Centre Léon Bérard, F-69373 Lyon, France

E-mail: david.sarrut@creatis.insa-lyon.fr

Received 15 January 2013, in final form 19 March 2013
Published 14 June 2013
Online at stacks.iop.org/PMB/58/4563

Abstract

Online dose monitoring in proton therapy is currently being investigated with prompt-gamma (PG) devices. PG emission was shown to be correlated with dose deposition. This relationship is mostly unknown under real conditions. We propose a machine learning approach based on simulations to create optimized treatment-specific classifiers that detect discrepancies between planned and delivered dose. Simulations were performed with the Monte-Carlo platform Gate/Geant4 for a spot-scanning proton therapy treatment and a PG camera prototype currently under investigation. The method first builds a learning set of perturbed situations corresponding to a range of patient translation. This set is then used to train a combined classifier using distal falloff and registered correlation measures. Classifier performances were evaluated using receiver operating characteristic curves and maximum associated specificity and sensitivity. A leave-one-out study showed that it is possible to detect discrepancies of 5 mm with specificity and sensitivity of 85% whereas using only distal falloff decreases the sensitivity down to 77% on the same data set. The proposed method could help to evaluate performance and to optimize the design of PG monitoring devices. It is generic: other learning sets of deviations, other measures and other types of classifiers could be studied to potentially reach better performance. At the moment, the main limitation lies in the computation time needed to perform the simulations.

(Some figures may appear in colour only in the online journal)

1. Introduction

In the last few years, proton therapy became an increasingly used modality for cancer treatment with more than 83 000 treated patients over the last 30 years worldwide (Durante and Loeffler 2009, PTCOG 2012). The proton depth-dose profile, with the so-called Bragg peak, leads to improved dose conformation to the tumor, compared with state-of-the-art intensity modulated radiation therapy (Lomax 1999, Mock *et al* 2004). Proton therapy allows tumors to be exposed to higher doses while limiting the energy deposited in healthy surrounding tissues (Smith 2009, Chera *et al* 2009), particularly after the distal falloff. However, the full potential of the proton ballistics cannot be fully exploited yet as uncertainties remain in the proton range, that could in some cases reach up to 5–15%, 5–6 mm according to (Andreo 2009, Smith 2009, Paganetti 2012). Clinicians generally avoid placing organs at risk behind the Bragg peak. Range uncertainties are notably due to the stoichiometric calibration of the planning CT scan, to organ motion, to inter-session anatomical changes and to patient mispositioning (Paganetti 2012).

Protons undergoing nuclear reactions with target nuclei create radioisotopes as well as high-energy secondary particles. As there is no transmission of the primary beam through the patient as is the case with x-rays, the secondary radiation going out of the patient is the only direct source of information to monitor the treatment. It has been shown that the spatial distribution of the secondary particle production is correlated with the dose distribution and the ion range inside the patient (Parodi and Enghardt 2000, Testa *et al* 2008). It was first proposed to exploit the annihilation gamma-rays originating from positron emitters (^{11}C, ^{15}O and others) produced by nuclear interaction along the beam path. Conventional PET imaging can be used but counting statistics has to be accumulated for as long as 2–30 min due to the rather low activity and positron emitters half lives ($\sim$20 min for ^{11}C and $\sim$2 min for ^{15}O) (Parodi *et al* 2002). During this time, patient motion and biological washout occur, which intrinsically limits PET monitoring (Attanasi *et al* 2011, España *et al* 2011, Moteabbed *et al* 2011). Developments are in progress to try to circumvent these issues by using time of flight (TOF) techniques (Karp *et al* 2008).

Prompt radiation monitoring is another option investigated to overcome the above-mentioned limitations. Depending on the particle used in the incident beam (proton, carbon ion, etc), the use of different types of secondary particles have been studied: (i) Prompt gamma (PG) (Stichelbaut and Jongen 2003, Min *et al* 2006, Testa *et al* 2008) and (ii) secondary protons with the so-called interaction vertex imaging (IVI) (in carbon-ion therapy) (Henriquet *et al* 2012). The present paper focuses on PG imaging only, although the proposed method could probably be applied to IVI as well.

PG monitoring is also being studied to overcome PET monitoring limitations. PG are photons created by inelastic interactions between incident proton and target nuclei. Unlike annihilation photons, PG are emitted quasi-instantaneously (decay time much smaller than 1 ps), with a very broad energy spectrum (from a few 10^5 eV to a few 10^7 eV). Most of them have enough energy to escape the patient. Typically for a $480 \times 480 \times 234$ mm^3 water phantom and a 182 MeV proton beam, we observed that 80% of PG escape the target, whereas only 48% of generated gamma do (49% for neutron related gamma, 26% for bremsstrahlung, 64% for positron annihilation). Their use for treatment monitoring faces two main issues: (i) efficiency issues due to their high energy and (ii) discrimination of the PG signal from an intense background noise caused by secondary neutrons.

Testa *et al* propose a PG collimated camera design with multiple slits perpendicular to the beam axis and scintillating crystals coupled to photomultiplier tubes (Testa *et al* 2008). Using a beam-tagging device (hodoscope), one can estimate the position of the emission of a detected

PG as the intersection of the beam axis and the slit plane. TOF and energy windows can be used to discriminate prompt PG from neutrons, thus improving the signal-to-background ratio. Other collimator geometries such as knife-edge slits are investigated by several research groups (Bom *et al* 2012, Jongen and Stichelbaut 2009, Smeets *et al* 2012). Richard *et al* (Richard and Chevallier 2010, Roellinghoff *et al* 2011, Richard 2012) proposed a Compton camera with a scatterer consisting of a stack of silicon strip detectors and a position-sensitive scintillating absorber. Events consisting of a Compton interaction in the silicon detector and subsequent absorption of the scattered photon in the scintillator make it possible to reconstruct the emission point of the PG as the intersection of a cone (using the Compton kinematics) and beam axis (Frandes *et al* 2010, Richard and Chevallier 2010). Spatial resolution was estimated by Monte-Carlo simulations to be about 7 mm full width at half maximum with a detection efficiency of 3×10^{-4} (Richard 2012). In this paper, we considered as a test case a multi-slit collimated camera coupled with a hodoscope, however the proposed method can be applied to other types of camera.

Two principal proton beam delivery techniques have been investigated under clinical conditions: passive spreading and active delivery systems. For passive delivery, spread out Bragg peak are formed by superimposition of shifted pristine Bragg peaks using range modulation wheel. For active delivery, magnets deflect and steer pencil beams (Lomax *et al* 2004) that are delivered by successive layers of decreasing energy. According to the production yields and an estimation of the camera detection efficiency (Testa *et al* 2008, Roellinghoff *et al* 2011), this delivery technique could potentially allow dose monitoring for the whole treatment, for a given energy layer, or even for a single spot. In this study, we investigated spot-by-spot monitoring.

PG dose monitoring aims at detecting deviations from TPS by using measured PG depth profiles and reference data from TPS. Those deviations impair dose delivery and change characteristics of detected PG profile. Several authors (Min *et al* 2006, Moteabbed *et al* 2011, Testa *et al* 2008) proposed to use PG profile falloffs to detect them. Indeed, a change of density or a shift along the beam path generally results in a shift of the PG profile falloff and can be detected with the camera. However, some deviations have no influence on falloff positions (see section 4). Therefore, we propose using other measures as well, such as the registered correlation of PG profiles. We also propose a machine learning methodology based on simulations to build classifiers during the planning stage that could then be used during a treatment session. The classifiers are specific to the treatment plan of a given patient and can potentially detect more deviations compared with the use of distal falloff alone. The method is generic: other measures and types of classifiers can be used. Kuess *et al* proposed a similar approach using in-beam PET data, using activity map as direct input of the machine learning algorithm, deviations are purely in-beam as the authors add material between the nozzle and the patient to change the beam penetration depth (Kuess *et al* 2012).

This paper is organized as follows. The section 2 presents simulations of a realistic setup combining proton delivery and monitoring. The section 3 presents the proposed machine learning approach to detect discrepancies between planned and delivered treatment. The proposed framework is evaluated in section 4 with receiver operating characteristics (ROC) curves and a leave-one-out (LOO) study. Advantages and limitations are discussed in section 5.

2. Simulations of treatment and dose monitoring

We describe in this section the simulation setup used by the machine learning approach described in section 3. The setup is intended to be as realistic as possible. It includes a

spot-scanning proton treatment plan of a prostate cancer patient and the complete description of a two-head collimated PG camera around the patient.

2.1. Treatment plan

We considered a prostate cancer treatment plan created with XIO TPS (Elekta). The stoichiometric calibration of the patient's CT image was performed as in Schneider *et al* (2000). The plan was composed of two laterally opposed fields, 2300 spots and 25 energy layers from 143 to 187 MeV. This corresponds to a conventional treatment of 2 Gy by fraction in the target area (PTV), 80 Gy in total. Each spot is described as proposed in Grevillot *et al* (2011): the optical and energy parameters of the beams were modeled using measured depth-dose profiles and spot sizes obtained from a clinical facility. As the number of protons in each spot (pencil beam) is given in Monitor Units in the treatment plan, we made a first low statistic simulation (5×10^6 protons) of the whole plan to determine the link between the dose in the PTV (1.7×10^{-4} Gy) and the number of incident protons. In general, distal spots are associated with larger weights than the ones of proximal spots. In the following we consider a single spot going through the center of the prostate with the Bragg peak near the distal part of the prostate. This spot has 50×10^6 protons and an energy of 182 MeV.

2.2. PG monitoring system

The online dose monitoring device simulated in this study is an extrapolation of a current PG camera prototype being investigated in Testa (2010). It is a two-head cylindrical collimated multi-slit detector. Each head is composed of 91 tungsten septa with thickness of 2 mm, interleaved with 2 mm air gaps, forming a 360 mm field of view. The collimation length is 100 mm, the distance from collimator to axis is 300 mm and detectors are placed 600 mm away from the beam axis to allow TOF measurement. The collimators spanned 50° around the beam, allowing the patient to rest on the treatment table. The detectors are BGO scintillators located after the collimation blades. They span 40° for each head and have a thickness of 50 mm. The spacing between the collimator and the BGO crystal is such that secondary radiations emitted by the collimators and impinging on the crystal are minimum (i.e. half-way which is a matter of solid angles). A beam-tagging device (hodoscope) is used to perform TOF filtering. This setup has roughly the same dimensions as the PET head used in the literature ($400 \times 360 \times 375$ mm) (Moteabbed *et al* 2011) and is illustrated in figure 1.

The PG camera is used to estimate the positions of the PG emission points inside the patient. Using the beam structure measured with the hodoscope, one can trigger the TOF window around the time PG are expected to arrive. When the TOF window closes, if the integrated energy deposited in a crystal lies in the acceptable energy window, the event is recorded. The position of the event in the crystal is considered as the energy weighed barycenter of all interactions in the crystal, plus a random value taken from a 5 mm FMHM Gaussian noise to simulate the electronics and the detector resolution (Richard and Chevallier 2010). More advanced model could be used but we observed that this noise is small compared to the one due to low statistic. Hence, the reconstructed PG profiles are histograms filled with event positions. Following (Testa 2010), events are selected if the deposited energy is above 2 MeV and TOF is in the range [3–6 ns] after the proton impacts in the hodoscope.

2.3. Monte-Carlo simulations

Simulations were carried out with GATE, a Geant4-based Monte-Carlo code (Jan *et al* 2011). We used GATE V6 and Geant4 V9.4p01. We used the physics list proposed in (Grevillot *et al*

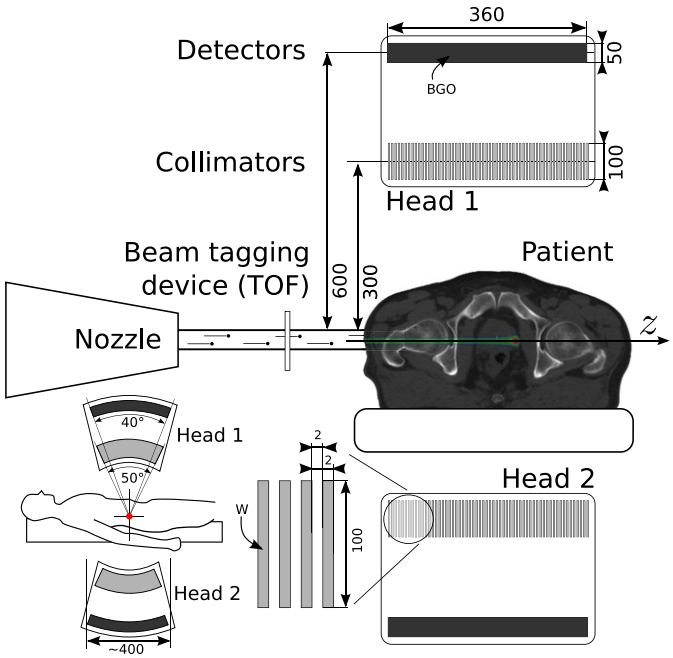


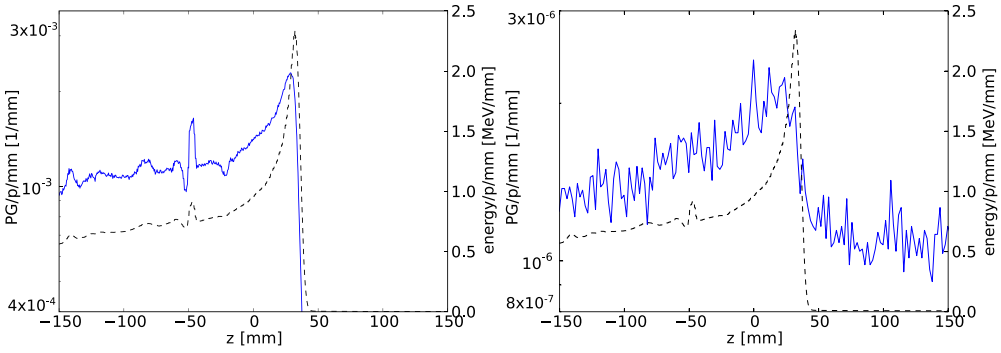
Figure 1. Simulation setup with schematic description of the PG two-head camera. Protons coming from the nozzle trigger the beam-tagging device and enter the patient on the patient's right side to form a Bragg peak in the patient's prostate. PG are collimated by the tungsten blades (in light gray), located in two heads above and below the patient. Dark gray represents the BGO crystals that detect gamma rays in an Anger-like fashion. Each head measures $360 \times 300 \times 400\,\text{mm}^3$, similar to the PET head used in Lomax *et al* (2004). Sizes are given in millimeters. The z-axis is the spatial axis along the camera field of view.

2011). Simulated and measured PG yields were shown to match around 10% in the case of protons (Polf *et al* 2009) and at a factor of 2 in the case of carbon ions (Dedes *et al* 2012), but we do not study carbon beams here.

Due to the lack of room geometry and beam structure simulation, background neutron noise is lower than in experimental data. However, this does not influence significantly detected PG profiles since most of the noise is rejected by TOF filtering. As simulation precision is not the main purpose of this paper, we consider it is sufficient to show the feasibility of the proposed methodology. The simulation of the spot we chose took about ten days on a single CPU 2.6 GHz Intel Xeon (60 p s^{-1}). We used the GateLab system (Camarasu-Pop *et al* 2010) to reduce this time to about 5.2 h (average speed up of 45, including queuing time and the merging of partial results). The GateLab is an open-source system that allows to submit Gate simulations on a large computing infrastructure such as the EGI grid from a simple web page. It can be used from www.opengatecollaboration.org/GateLab. It took less than five days to complete all simulations needed to apply the machine learning approach described in section 3. Note that no particular effort to reduce the computation time has been performed.

2.4. Observables

In addition to PG profiles, simulations also stored the deposited energy and the emission PG profiles as 1D distributions inside the patient, along the proton beam and perpendicular



(a) PG emission (b) detected PG

Figure 2. For the pencil beam we chose, (a) PG emission profile (left axis, plain curve) and (b) detected PG profile (left axis, plain curve). The depth-dose profile is shown along with the two profiles (right axis, dashed curve). The histogram bins are 0.45 mm wide for energy deposition and PG emission profiles and 2 mm wide for detected PG profiles. Events are selected by the camera if the deposited energy is above 2 MeV and TOF in the range [3–6 ns]. The origin of the z-axis corresponds to the center of the camera.

to the PG camera collimators. Figure 2 illustrates the output of the simulation for a given treatment spot. As shown in several studies (e.g. Moteabbed *et al* 2011), one can observe that the PG falloff and the Bragg peak are distant from few millimeters. PG emission rate is around 10^{-3} PG/p/mm, increases rapidly before the Bragg Peak and decreases dramatically after. The counting rate of the detected profile is about 10^{-6} count/p/mm, with high level of statistical noise and a reduced contrast around the Bragg peak. 1 out of 1000 PG is detected in the camera. Detectors should be segmented to avoid the saturation of individual photomultiplier.

3. A machine learning approach

3.1. Principle

Given a training data set of known situations, composed of a set of input cases X_i (i.e. PG detected profiles) with corresponding output Y_i (e.g. patient displacements), a machine learning algorithm aims at building a function F able to infer outputs from inputs. If the output is discrete the function is called a classifier, and if it is continuous, it is called a regression function. This process is called supervised learning in the sense that output is known in the training set (TS). The function F, once tuned to the relationship between input and output, can be used to predict the correct output for any input. The main objective and difficulty of this stage is the generalization: the ability to accurately predict correct output from input that does not belong to the initial training.

We intend here to investigate the contribution of such a concept to the issue of detecting deviations using PG monitoring. We define a TS, build a regression function based on practical considerations and test the generalization with a LOO approach. Numerous methods exist in the literature (Kotsiantis *et al* 2007), such as decision trees, neural networks, genetic programming, support vector machines (SVM), Bayesian networks, etc. We decide here to focus on a simple threshold-based approach, based on ROC curves, but more advanced methods could be investigated.

3.2. Training set

To build the TS $\{(X_i, Y_i)\}$, we considered deviations in the form of patient translations. Other types of deviations could be studied, such as patient rotations, anatomical changes or errors in the delivery of the planned beam. Each element of the TS corresponds to a difference between two simulations S_j and $S_{j'}$. The output Y_i is the patient translation between S_j and S'_j, and the input X_i is a measure of a distance between detected PG profiles, described in the next section. Alternatively one could use gamma-index ratio for Y_i. We simulated 20 patient positions, $S_1, \ldots, S_{20}$, and considered unordered pairs of situations $((j, j') = (j', j))$, without self reference $(j \neq j')$. We thus obtain $(20 \times 19)/2 = 190$ elements in the TS. This approach provides a TS with a reasonable size and a limited number of simulations.

Patient positions were generated in order to get a uniform distribution of Y_i in-beam and off-beam components in the range [2–22 mm]. This ensures that the machine learning method is not biased by an unbalanced representation of certain events in the TS. As we consider unordered pairs of positions, uniform distribution cannot be easily generated: to reach that goal, we randomly generated candidate positions until the distribution was uniform. It was performed by an optimization process that minimizes the distance between the target uniform distribution and the in-beam and off-beam histograms built from the randomly chosen positions. We used the Nelder–Mead optimization algorithm (Nelder and Mead 1965). The lower bound of the range [2–22 mm] has been chosen because too low distances lead to poorer classifier performances. The upper bound was chosen as a reasonable upper displacement value. Further studies on the TS definition are needed but are beyond the scope of this paper.

3.3. Distance measures on PG profiles

Having defined Y_i, we now need to define X_i. The simulated deviations between situations S_j and $S_{j'}$ will induce changes in the PG profiles that may be quantified with measures between the two profiles. If the patient is translated along the beam axis (in-beam deviation), the Bragg peak will be shifted but the profile will roughly keep its shape. On the other hand, if the displacement is perpendicular to the beam axis (off-beam deviation), the falloff will not be shifted significantly, however the shape of the profile will probably be affected. We thus investigate two measures: distal falloff position difference and registered correlation. Distal falloff position difference has already been used in the literature (Parodi *et al* 2007) but is known to fail predicting off-beam deviation components (Bom *et al* 2012). The proposed second measure, registered correlation, has been introduced to overcome distal falloff limitations and predict off-beam deviations.

The distal falloff position is measured using the following algorithm (see figure 3). A quadratic spline is fitted on a 100 mm Bragg peak window centered at the maximum amplitude (value) of the PG profile. The minimum amplitude is the mean PG value on the tail window, located after the Bragg peak region. The 50% distal falloff position, represented by the range (S_j) operator, is the position of the point located on the spline that has an amplitude equal to the mean value between the maximum and the minimum amplitudes. The measure $\Delta_{\text{range}}(S_j, S_{j'})$ is defined as the absolute difference between the range values of the reference and observed PG profiles as in equation (1).

$$\Delta_{\text{range}}(S_j, S_{j'}) = \|\text{range}(S_j) - \text{range}(S_{j'})\| \tag{1}$$

The second measure, registered correlation, is built to be both correlated with off-beam deviation and not correlated with in-beam deviation. The decorrelation with the in-beam component of the deviation ensures an easy interpretation of the measure and facilitates the training process of the combined classifier defined in 3.4. Ideally, registered correlation must

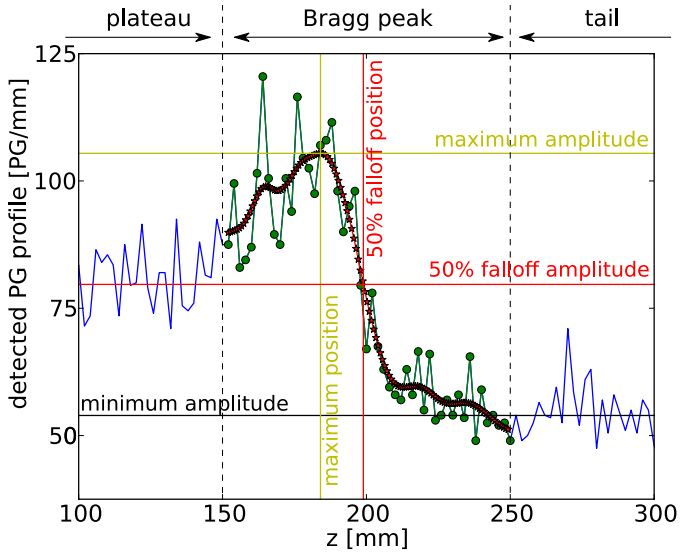


Figure 3. Measure of the distal falloff position on the detected profile of figure 2(b). The blue curve is the detected PG profile. It is split into three windows with fixed size, positioned relative to the maximum amplitude of the profile. The green dots depict the Bragg peak window used to create the spline. The red curve is the corresponding fitted spline. The tail window is used to compute the minimum amplitude. The plateau region is used while calculating the registered correlation measure.

be taken between profiles expressed in the patient coordinate system to cancel the in-beam deviation influence on the measure. To decorrelate with in-beam deviation, the estimated falloff position, range(S_j), is used since it proved to be much more robust with fitted spline than with the maximum of correlation. Once the registration is performed (with curves shift), the correlation coefficient is computed on windows taken from the plateau part (before the Bragg peak) of PG profiles. Registering the images ensures a greater decorrelation with the in-beam deviation since a small residual shift between Bragg peak positions has a large effect on the correlation value. For 182 MeV protons, plateau windows are 120 mm wide and end 50 mm before the Bragg peak. Their upper bound is the same as the lower bound of the spline used to define range(S_j) (see figure 3). Equation (2) defines the registered correlation operator $\mathrm{corr}_{\mathrm{reg}}(S_j, S_{j'})$, where $X_j(z)$ (resp. $X_{j'}(z)$) is the PG profile of simulation S_j (resp. $S_{j'}$). z is a continuous spatial coordinate along the camera axis as defined in figure 1. $X_j(z + \mathrm{range}(S_j))$ and $X_{j'}(z + \mathrm{range}(S_{j'}))$ are the registered PG profiles so that $z = 0$ corresponds to the PG profile falloff for both profiles.

$$\mathrm{corr}_{\mathrm{reg}}(S_j, S_{j'}) = \mathrm{corr}(X_j(z + \mathrm{range}(S_j)), X_{j'}(z + \mathrm{range}(S_{j'}))) \tag{2}$$

corr is the correlation operator restricted to the plateau window and is defined by equation (3).

$$\mathrm{corr}(X, Y) = \int_{z=-170\,\mathrm{mm}}^{z=-50\,\mathrm{mm}} (X(z) - \bar{X})(Y(z) - \bar{Y})\,\mathrm{d}z \tag{3}$$

$\bar{X}$ is the mean value operator with support limited to the plateau window and is defined by : $\bar{X} = \int_{z=-170\,\mathrm{mm}}^{z=-50\,\mathrm{mm}} X(z)\,\mathrm{d}z$.

Other off-beam detection measures can be defined as long as they measure changes in profile shape in the patient frame of reference. For example, we could have used the sum of

absolute differences between $X_j(z - \text{range}(S_j))$ and $X_{j'}(z - \text{range}(S_{j'}))$ instead of correlation, but the initial results were less good. As the purpose of this paper is not to provide the best possible discriminant observable, but rather to study the overall procedure, we did not further investigate other variables

3.4. Building the classifiers

Once the TS is defined, we train a classifier that predicts the displacement amplitude from the measures on PG profiles. We propose to use threshold classifiers. This type of classifiers uses a tolerance parameter that we call here deviation threshold (DT), defined by the user. For simplicity, we considered here the maximum tolerated patient translation. More generally, another tolerance could be defined, such as a value related to the gamma-index ratio between the two compared dose distributions.

Training a classifier is the process of finding the measurement threshold (MT) that gives the best performance. It must be repeated for each treatment plan. Using the TS and a fixed MT, one can determine the cardinality of the four following cases: true positive (TP) when deviation and measure are greater than, respectively, DT and MT; true negative (TN) when deviation and measure are below both thresholds; false positive (FP) when something is detected (measure greater than MT), but the deviation is still within tolerance (deviation below DT) and false negative (FN) when the treatment goes wrong (deviation greater than DT) but nothing is detected by the measure (measure below MT). FN represents the worst case, because it means that the patient will undergo dose discrepancies, without these being detected by the system. One would also want to reduce FP, because they trigger false alarm and reduce the patient throughput of the treatment machine. Based on the four cases, Equation (4) defines the true negative rate (TNR, also known as specificity) and the true positive rate (TPR, also known as sensitivity).

$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}} \qquad \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

When training the classifier, one wants to maximize both TNR and TPR to minimize FP and FN. The ROC curve displays TPR versus 1-TNR for a range of MT and a fixed DT. We thus define the associated specificity and sensitivity (ASS), as in equation (5), to be maximized during the training process. This is equivalent to finding the most upper/left point on an ROC curve. It is called maximum associated specificity and sensitivity (MASS) (Waghorn *et al* 2011) and its value is a classifier performance measure. The associated MT is the optimal threshold that gives the best classification performance.

$$\text{ASS} = \sqrt{\text{TNR}^2 + \text{TPR}^2} \qquad \text{MASS} = \max_{\text{MT}} \text{ASS} \tag{5}$$

4. Results

Experiments were conducted for the TS composed of 190 elements as described below, with three classifiers. The first two ones use the proposed measures separately: C_{FO} uses Δ_{range} and C_{RC} uses corr_{reg}. The last one C_{comb} combines both measures. It is triggered when an in-beam component, predicted by Δ_{range}, *or* an off-beam component, predicted by corr_{reg}, are higher than DT. To train it, one needs to find two MT, one for Δ_{range} and one for corr_{reg}, that give the best ASS. This is done by performing an exhaustive search on the measurement space.

We evaluated the prediction ability of the classifiers by a LOO procedure. This method allows assessment of how the results of a classifier will generalize to data independent from the TS. One element is removed from the TS. The 189 remaining elements are used to train

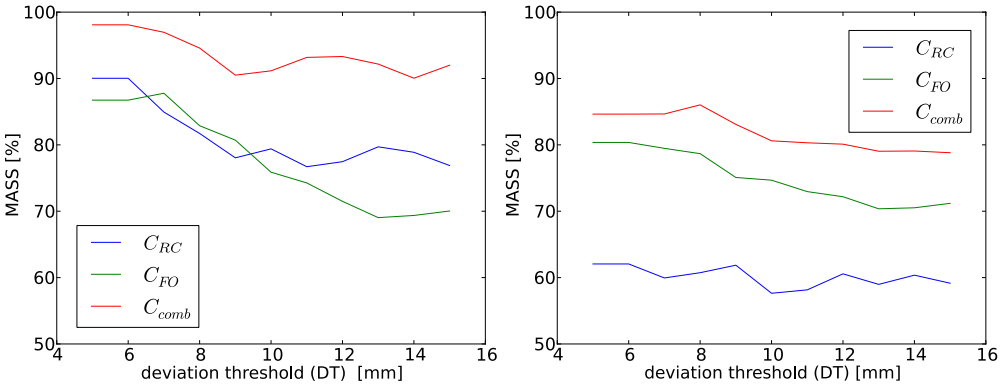


Figure 4. Performance (MASS) of C_{FO}, C_{RC} and C_{comb} according to different DT values. The left figure corresponds to case A and the right figure to case B.

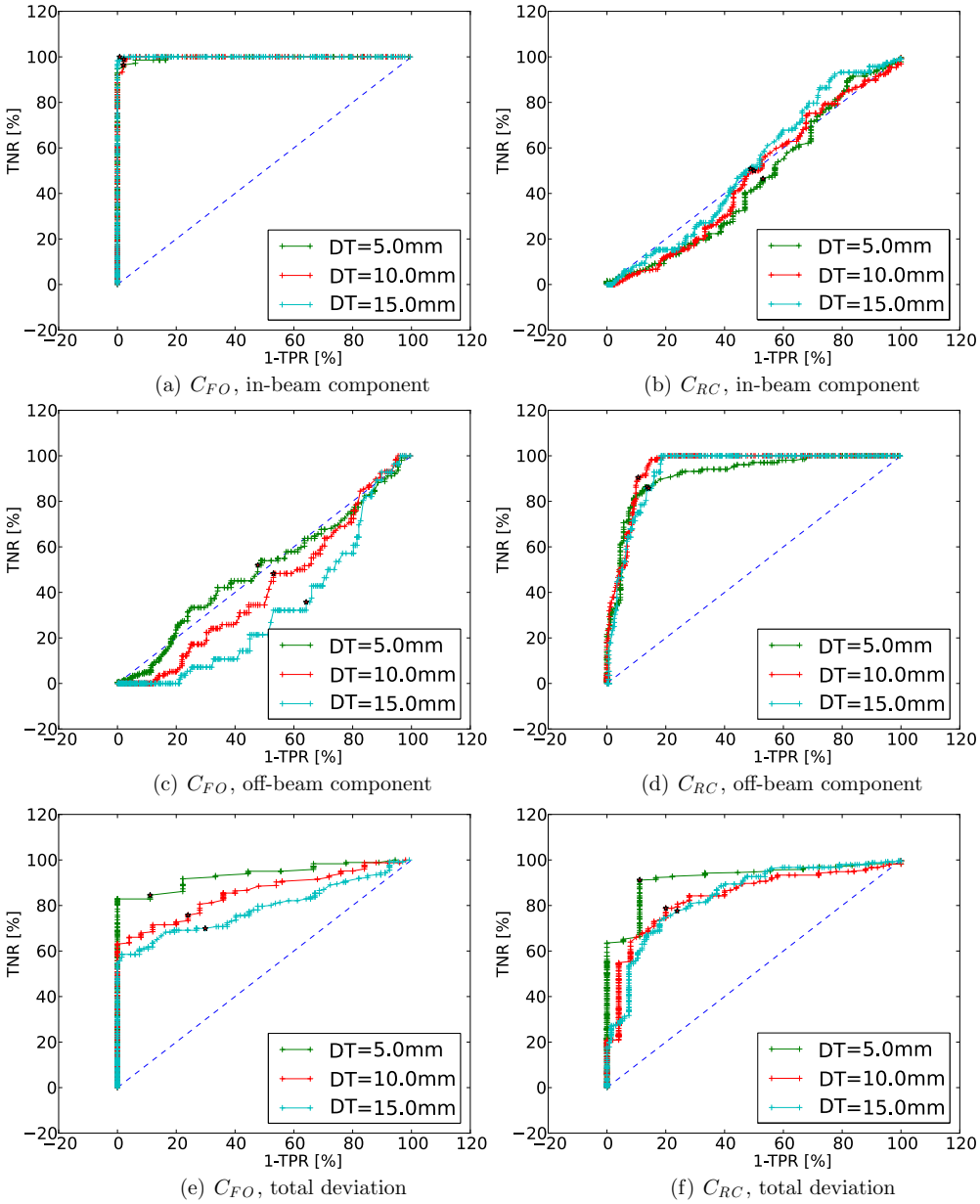
Table 1. Performance of the three classifiers C_{FO}, C_{RC} and C_{comb} evaluated with the LOO method, for cases A and B. DT was set to 5 mm.

	TP	FN	TN	FP	TNR	TPR	MASS
Case A: production yields (ideal case)							
C_{RC}	86.8%	8.4%	4.2%	0.5%	91.2%	88.9%	90.0%
C_{FO}	80.5%	14.7%	4.2%	0.5%	84.5%	88.9%	86.7%
C_{comb}	91.6%	0.0%	4.7%	3.7%	96.1%	100%	98.1%
Case B: PG profiles measured with the camera							
C_{RC}	60.0%	35.3%	2.9%	1.8%	63.0%	61.1%	62.1%
C_{FO}	78.9%	16.3%	3.7%	1.1%	82.9%	77.8%	80.4%
C_{comb}	81.8%	0.8%	3.9%	13.4%	85.9%	83.3%	84.6%

classifiers, which then predict the class of the removed element. This evaluation is performed 190 times, by successively removing all elements, and the results are averaged. This evaluates the performance and the robustness of the classifier. Finally, the procedure is repeated two times: one time with PG emission profiles, which correspond to an 'ideal case', referred to as case A, and one time with the PG profiles detected by the simulated camera, much closer to a real situation, referred to as case B.

Table 1 presents the LOO prediction results (TPR, TNR and MASS) obtained with the three classifiers for DT = 5 mm. Figure 4 shows the influence of the DT on the performance of the classifier (MASS) for cases A and B. Figures 5 and 6 depict ROC curves obtained from the TS by decomposing the deviation into in-beam and off-beam components. Three values of DT were used: 5, 10 and 15 mm. This range is chosen to ensure that each class is significantly populated.

We observed in table 1 and figure 4 that the combined classifier leads to better results compared with the others, for both cases. It is possible to detect discrepancies of 5 mm with TNR and TPR of around 85%, whereas using only distal falloff leads to 80% on the same TS. C_{RC} leads to the poorest results because it is designed to detect features other than the falloff differences. The TPR of C_{comb} increases when compared with C_{FO} alone. It shows that the combination of the two measures helps to detect deviations having combined in-beam and off-beam components. This is particularly the case for case A and less pronounced with case B. Case A should be interpreted as a reference to strive toward. As shown in figure 2,



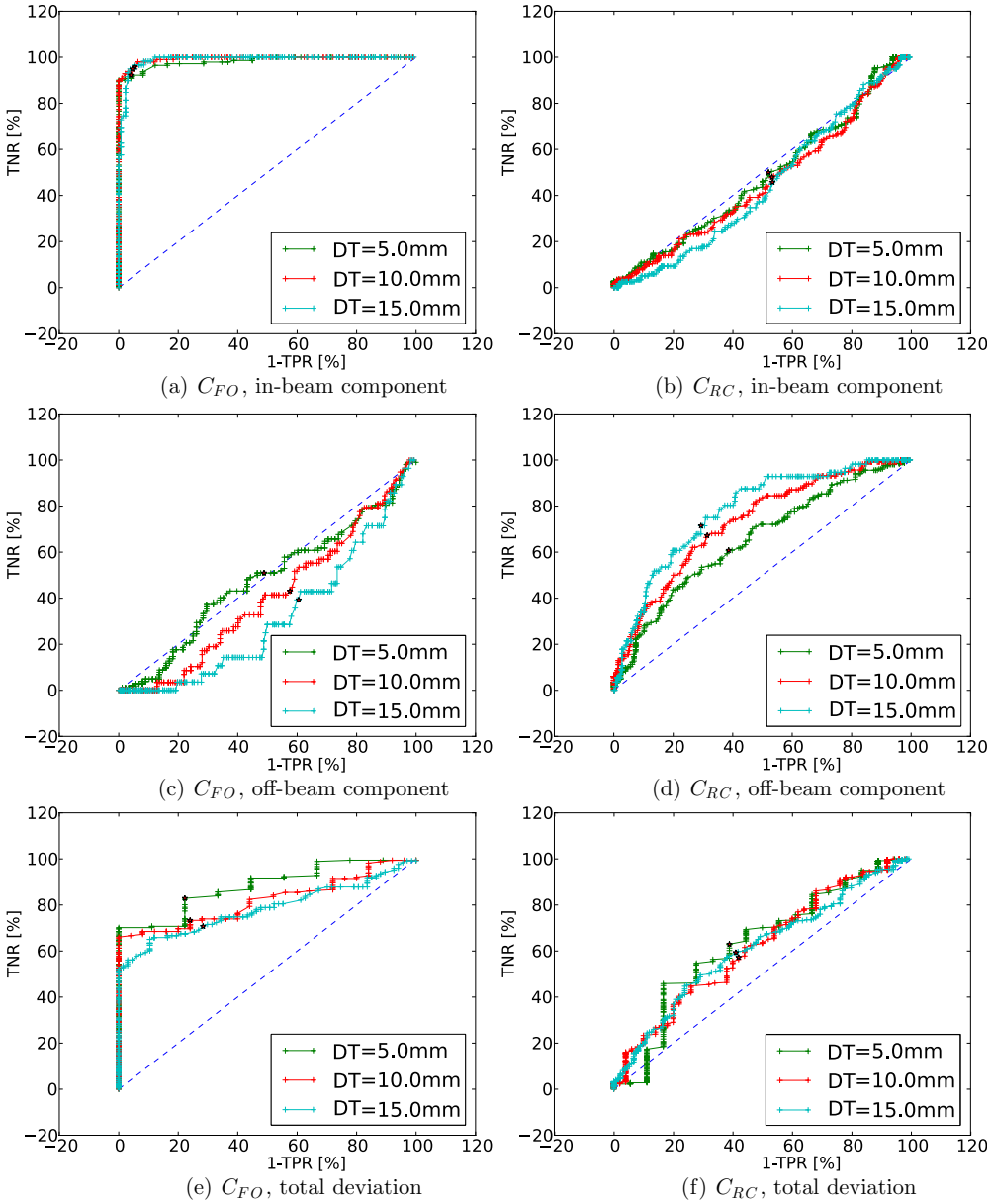
(a) C_{FO}, in-beam component (b) C_{RC}, in-beam component

(c) C_{FO}, off-beam component (d) C_{RC}, off-beam component

(e) C_{FO}, total deviation (f) C_{RC}, total deviation

Figure 5. ROC curves for the two classifiers C_{FO} (first column) and C_{RC} (second column), according to in-beam (first row) and off-beam (second row) components of the deviation, and the total deviation (third row). Corresponding DT are given in the legend. MASS points are represented by stars. Classifiers were trained with PG emission profiles, case A.

the statistics are three orders of magnitude lower in case B than in A (10^{-3} count/p/mm to 10^{-6} count/p/mm). An improved camera design or better statistics using several spots instead of a single one, could help to improve prediction.



(a) C_{FO}, in-beam component (b) C_{RC}, in-beam component

(c) C_{FO}, off-beam component (d) C_{RC}, off-beam component

(e) C_{FO}, total deviation (f) C_{RC}, total deviation

Figure 6. Same figure as 5 but classifiers were trained with detected PG profiles, case B.

Figure 4 shows that the DT has a relatively low influence on the performance in the range [5–15 mm]. The combined classifier shows a great interest in the ideal case A, while the increased noise in the detected PG profiles leads to only slightly better results than using C_{FO} alone for case B. Note as well that performance is degrading while DT is increased. This may seem counter-intuitive, but this is due to the fact that above a certain deviation, profiles tend to decorrelate completely from one another and classifiers cannot make the difference between

highly and very highly perturbed setups. For DT < 6 mm in case A or DT < 8 mm in case B, performance stabilizes because of the intrinsic noise of the profile.

Finally, the ROC curves in figures 5 and 6 provide an insight on the performance of the classifiers. On these curves, poor performance corresponds to a MASS value close to 50% and the corresponding ROC curve is close to $x = y$. On the other hand, classifiers having good performance lead to a higher MASS value and, ideally, should tend toward a value near the upper left corner of the graph. The curves illustrate that the two measures try, by design, to capture the two main components of the deviation, the in-beam and the off-beam parts. In the ideal case A, C_{FO} almost perfectly predicts the in-beam component (TPR and TNR close to 100%), while it is close to a random prediction (50%) for the off-beam component. Predicting the total deviation, as happens in practical situations, leads to decreased performance, in particular in terms of TPR. Conversely, C_{RC} fails for in-beam component but leads to interesting performance for the off-beam one. Of course, it is much more difficult to detect such deviations and the performance is lower than C_{FO} with the in-beam component. In case B, performance largely decreases due to the low statistics (TPR and TNR around 60–70%). The C_{RC} measure seems to be more degraded than C_{FO}. However, as currently there is no dose validation during proton therapy, obtaining 84% is already quite promising, and in any case better than nothing.

5. Discussion and conclusion

The proposed method investigates the potential of a PG-based dose monitoring device in clinical conditions by a machine learning approach. To our knowledge, no PG camera has been used in a clinical situation, so we used simulations that combine dose deposition in a patient CT image and profiles measured with a PG camera. The main test case characteristics were the following: (1) the whole PG camera design was an extension of the prototype proposed in Testa *et al* (2008), still in development, (2) a realistic prostate treatment plan was considered with a patient CT description, (3) the considered treatment deviations were translations only, along the beam direction and in the transverse plane (4) a simple classifier was used, with two proposed measures: difference of falloff positions and correlation between the registered profile plateau regions.

Building a classifier highly depends on the TS but it is unclear how to optimally build a representative and minimal size TS. Of course, it is not realistic to simulate all deviations that can potentially occur during a treatment together with all spots in the treatment plan. The proposition here was to consider a single spot and a set of random translations equally distributed in the predefined range [2–22 mm]. By doing this we cover a certain part of the deviation space with a minimal amount of simulations. Further studies should be performed to investigate the impact of the construction of the learning space. Moreover, deviations of types other than translations (e.g. rotations, errors in stoichiometric calibration, anatomical changes) could be included in the TS. Then, one should use patient dose gamma index ratio as a consolidated input for classifiers. DT would then be the minimum gamma index ratio tolerable during patient treatment. Gamma index would be computed on dose distributions obtained from the simulations. Also, this method does not distinguish the origins of the deviations (translations, anatomy changes, ...) but only their consequences on the PG profiles, according to a tolerance threshold.

We observed that the number of PG detected by the camera (case B) should be improved when observing a single spot, whereas case A works fine. We can expect a substantial increase of the camera efficiency thanks to further optimization in progress. Besides, a combined

classifier could be trained with multiple spots to reach better statistics and increase classifier performances.

C_{FO} and C_{RC} classifiers were combined with a 'or' operator to give priority to a low number of FN. Other types of combinations could be proposed together with other types of classification methods, such as the SVM, with potentially higher generalization capabilities.

As a conclusion, we think that the proposed method could help to evaluate the performance of PG monitoring devices and to improve their design. It is generic: other TS, other measures and other types of classifiers could be studied to potentially reach better performance. It could also potentially be applied to other types of monitoring technologies, such as Hadron-PET or IVI (Henriquet *et al* 2012), and provide a standard framework for comparing performance. Future clinical use will require simulating PG profiles with improved accuracy. The simulations, which are the most consuming part of the method, will have to be part of the treatment planning. Alternatives to Monte-Carlo methods could help reducing the computation time.

Acknowledgments

This work was supported in part by the European collaboration Envision (grant agreement no. 241851), the PRRH (Pogramme Régional de Recherche en Hadronthérapie) from the 'région Rhone-Alpes', and the Labex PRIMES (French ANR) and is part of the France Hadron research program.

References

Andreo P 2009 On the clinical spatial resolution achievable with protons and heavier charged particle radiotherapy beams *Phys. Med. Biol.* **54** N205–15

Attanasi F, Knopf A, Parodi K, Paganetti H, Bortfeld T, Rosso V and Del Guerra A 2011 Extension and validation of an analytical model for *in vivo* PET verification of proton therapy—a phantom and clinical study *Phys. Med. Biol.* **56** 5079–98

Bom V, Joulaeizadeh L and Beekman F 2012 Real-time prompt gamma monitoring in spot-scanning proton therapy using imaging through a knife-edge-shaped slit *Phys. Med. Biol.* **57** 297

Camarasu-Pop S, Glatard T, Mościcki J T, Benoit-Cattin H and Sarrut D 2010 Dynamic partitioning of GATE Monte-carlo simulations on EGEE *J. Grid Comput.* **8** 241–59

Chera B S, Rodriguez C, Morris C G, Louis D, Yeung D, Li Z and Mendenhall N P 2009 Dosimetric comparison of three different involved nodal irradiation techniques for stage II Hodgkin's lymphoma patients: conventional radiotherapy, intensity-modulated radiotherapy, and three-dimensional proton radiotherapy *Int. J. Radiat. Oncol. Biol. Phys.* **75** 1173–80

Dedes G *et al* 2012 Monte Carlo nuclear models evaluation and improvements for real-time prompt gamma-ray monitoring in proton and carbon therapy *Nuclear Science Symp., Medical Imaging Conf. & Workshop on Room-Temperature Semiconductor X-Ray and Gamma-Ray Detectors* at press

Durante M and Loeffler J S 2009 Charged particles in radiation oncology *Nature Rev. Clin. Oncol.* **7** 37–43

España S, Zhu X, Daartz J, El Fakhri G, Bortfeld T and Paganetti H 2011 The reliability of proton-nuclear interaction cross-section data to predict proton-induced PET images in proton therapy *Phys. Med. Biol.* **56** 2687

Frandes M, Zoglauer A, Maxim V and Prost R 2010 A tracking Compton-scattering imaging system for hadron therapy monitoring *IEEE Trans. Nucl. Sci.* **57** 144–50

Grevillot L, Bertrand D, Dessy F, Freud N and Sarrut D 2011 A Monte Carlo pencil beam scanning model for proton treatment plan simulation using GATE/GEANT4 *Phys. Med. Biol.* **56** 5203

Henriquet P *et al* 2012 Interaction vertex imaging (IVI) for carbon ion therapy monitoring: a feasibility study *Phys. Med. Biol.* **57** 4655–69

Jan S *et al* 2011 GATE V6: a major enhancement of the GATE simulation platform enabling modelling of CT and radiotherapy *Phys. Med. Biol.* **56** 881–901

Jongen Y and Stichelbaut F 2009 Device and method for particle therapy verification *US Patent Specification* 13/002484

Karp J S, Surti S, Daube-Witherspoon M E and Muehllehner G 2008 Benefit of time-of-flight in PET: experimental and clinical results *J. Nucl. Med.* **49** 462–70

Kotsiantis S B, Zaharakis I D and Pintelas P E 2007 Supervised machine learning: a review of classification techniques *Front. Artif. Intell. Appl.* **160** 3

Kuess P, Birkfellner W, Enghardt W, Helmbrecht S, Fiedler F and Georg D 2012 Using statistical measures for automated comparison of in-beam PET data *Med. Phys.* **39** 5874

Lomax A 1999 Intensity modulation methods for proton radiotherapy *Phys. Med. Biol.* **44** 185–205

Lomax A J *et al* 2004 Treatment planning and verification of proton therapy using spot scanning: initial experiences *Med. Phys.* **31** 3150

Min C H, Kim C H, Youn M Y and Kim J W 2006 Prompt gamma measurements for locating the dose falloff region in the proton therapy *Appl. Phys. Lett.* **89** 183517

Mock U *et al* 2004 Treatment planning comparison of conventional, 3D conformal, and intensity-modulated photon (IMRT) and proton therapy for paranasal sinus carcinoma *Int. J. Radiat. Oncol. Biol. Phys.* **58** 147

Moteabbed M, España S and Paganetti H 2011 Monte Carlo patient study on the comparison of prompt gamma and PET imaging for range verification in proton therapy *Phys. Med. Biol.* **56** 1063–82

Nelder J A and Mead R 1965 A simplex method for function minimization *Comput. J.* **7** 308–13

Paganetti H 2012 Range uncertainties in proton therapy and the role of Monte Carlo simulations *Phys. Med. Biol.* **57** R99–117

Parodi K and Enghardt W 2000 Potential application of PET in quality assurance of proton therapy *Phys. Med. Biol.* **45** N151–6

Parodi K, Enghardt W and Haberer T 2002 In-beam PET measurements of beta+ radioactivity induced by proton beams *Phys. Med. Biol.* **47** 21–36

Parodi K, Ferrari A, Sommerer F and Paganetti H 2007 A MC tool for CT-based calculations of dose delivery and activation in proton therapy *1st European Workshop on Monte-Carlo Treatment Planning* vol 74 (IOP Publishing) p 021013

Polf J C, Peterson S, McCleskey M, Roeder BT, Spiridon A, Beddar S and Trache L 2009 Measurement and calculation of characteristic prompt gamma ray spectra emitted during proton irradiation *Phys. Med. Biol.* **54** N519

PTCOG 2012 Patient statistics http://ptcog.web.psi.ch/patient_statistics.html. Accessed: 12/12/2012

Richard M H 2012 Conception d'une caméra Compton pour le contrôle en ligne en hadronthérapie *PhD Thesis* Université Claude Bernard-Lyon I

Richard M H and Chevallier M 2010 Design guidelines for a double scattering Compton camera for prompt-gamma imaging during ion beam therapy: a Monte Carlo simulation study *IEEE Trans. Nucl. Sci.* **58** 87–94

Roellinghoff F *et al* 2011 Design of a Compton camera for 3D prompt- imaging during ion beam therapy *Nucl. Instrum. Methods Phys. Res.* A **648** S20–23

Schneider W, Bortfeld T and Schlegel W 2000 Correlation between CT numbers and tissue parameters needed for Monte Carlo simulations of clinical dose distributions *Phys. Med. Biol.* **45** 459

Smeets J *et al* 2012 Prompt gamma imaging with a slit camera for real-time range control in proton therapy *Phys. Med. Biol.* **57** 3371–405

Smith A R 2009 Vision 20-20: proton therapy *Med. Phys.* **36** 556

Stichelbaut F and Jongen Y 2003 Verification of the proton beam position in the patient by the detection of prompt gamma-rays emission *Meeting of 39th Particle Therapy Co-Operative Group (San Francisco)*

Testa E, Bajard M, Chevallier M, Dauvergne D, Le Foulher F, Freud N, Létang J M, Poizat J C, Ray C and Testa M 2008 Monitoring the Bragg peak location of 73 MeV/*u* carbon ions by means of prompt gamma-ray measurements *Appl. Phys. Lett.* **93** 093506

Testa M 2010 Physical measurements for ion range verification in charged particle therapy *PhD Thesis* Université Claude Bernard Lyon 1

Waghorn B, Meeks S and Langen K 2011 Analyzing the impact of intrafraction motion: correlation of different dose metrics with changes in target $D_{95\%}$ *Med. Phys.* **38** 4505