"A noise correction of the γ-index method for Monte Carlo dose distribution comparison"

Cohilis, Marie ; Sterpin, Edmond ; Lee, John Aldo ; Souris, Kevin

ABSTRACT

Purpose Due to the increasing complexity of IMRT/IMPT treatments, quality assurance (QA) is essential to verify the quality of the dose distribution actually delivered. In this context, Monte Carlo (MC) simulations are more and more often used to verify the accuracy of the treatment planning system (TPS). The most common method of dose comparison is the γ-test, which combines dose difference and distance-to-agreement (DTA) criteria. However, this method is known to be dependent on the noise level in dose distributions. We propose here a method to correct the bias of the γ passing rate (GPR) induced by MC noise. Methods The GPR amplitude was studied as a function of the MC noise level. A model of this noise effect was mathematically derived. This model was then used to predict the time-consuming low-noise GPR by fitting multiple fast MC dose calculations. MC dose maps with a noise level between 2% and 20% were computed, and the GPR was predicted at a noise level of 0.3%. Due to the asymmetry of the γ-test, two different cases were considered: the MC dose was first set as reference dose, then as evaluated dose in the γ-test. Our method was applied on six proton therapy plans including analytical doses from the TPS or patient-specific QA measurements. Results An average absolute error of 4.31% was observed on the GPR computed for MC doses with 2% statistical noise. Our method was able to improve the accuracy of the gamma passing rate by up to 13%. The method was found especially efficient to correct the noise bias when the DTA criterion is low. Conclusion...

CITE THIS VERSION

Cohilis, Marie ; Sterpin, Edmond ; Lee, John Aldo ; Souris, Kevin. *A noise correction of the γ#index method for Monte Carlo dose distribution comparison.* In: *Medical Physics*, Vol. 47, no. 2, p. 681-692 (2020) http://hdl.handle.net/2078.1/226523 -- DOI : 10.1002/mp.13888

A noise correction of the γ-index method for Monte Carlo dose distribution comparison

Marie Cohilis[a)]
Center of Molecular Imaging, Radiotherapy and Oncology (MIRO), Institut de Recherche Expérimentale et Clinique (IREC), Université catholique de Louvain, 1200 Brussels, Belgium

Edmond Sterpin
Center of Molecular Imaging, Radiotherapy and Oncology (MIRO), Institut de Recherche Expérimentale et Clinique (IREC), Université catholique de Louvain, 1200 Brussels, Belgium
Department of Oncology, Laboratory of Experimental Radiotherapy, Katholieke Universiteit Leuven, 3000 Leuven, Belgium

John A. Lee, and Kevin Souris
Center of Molecular Imaging, Radiotherapy and Oncology (MIRO), Institut de Recherche Expérimentale et Clinique (IREC), Université catholique de Louvain, 1200 Brussels, Belgium

(Received 12 February 2019; revised 19 September 2019; accepted for publication 4 October 2019; published xx xxxx xxxx)

Purpose: Due to the increasing complexity of IMRT/IMPT treatments, quality assurance (QA) is essential to verify the quality of the dose distribution actually delivered. In this context, Monte Carlo (MC) simulations are more and more often used to verify the accuracy of the treatment planning system (TPS). The most common method of dose comparison is the γ-test, which combines dose difference and distance-to-agreement (DTA) criteria. However, this method is known to be dependent on the noise level in dose distributions. We propose here a method to correct the bias of the γ passing rate (GPR) induced by MC noise.

Methods: The GPR amplitude was studied as a function of the MC noise level. A model of this noise effect was mathematically derived. This model was then used to predict the time-consuming low-noise GPR by fitting multiple fast MC dose calculations. MC dose maps with a noise level between 2% and 20% were computed, and the GPR was predicted at a noise level of 0.3%. Due to the asymmetry of the γ-test, two different cases were considered: the MC dose was first set as reference dose, then as evaluated dose in the γ-test. Our method was applied on six proton therapy plans including analytical doses from the TPS or patient-specific QA measurements.

Results: An average absolute error of 4.31% was observed on the GPR computed for MC doses with 2% statistical noise. Our method was able to improve the accuracy of the gamma passing rate by up to 13%. The method was found especially efficient to correct the noise bias when the DTA criterion is low.

Conclusions: We propose a method to enhance the γ-evaluation of a treatment plan when there is noise in one of the compared distributions. The method allows, in a tractable time, to detect the cases for which a correction is necessary and can improve the accuracy of the resulting passing rates. [https://doi.org/10.1002/mp.13888]

Key words: γ-evaluation, Monte Carlo simulation, patient QA

1. INTRODUCTION

Radiotherapy techniques are in constant progress to improve the radiation delivery to the target, while allowing better sparing of healthy tissues. Intensity-modulated radiotherapy (IMRT) and, more recently, intensity-modulated proton therapy (IMPT) enable the delivery of very conformal dose distributions to the target.[1,2] This additional flexibility also comes with an increased complexity in the delivery equipment and treatment plan optimization algorithms. Therefore, quality assurance (QA) is essential to evaluate the quality of the treatment delivered in the patient.[3] In particular, patient-specific QA (PSQA) verifies that the delivered dose corresponds to the expected dose distribution computed by the treatment planning system (TPS). This procedure generally involves the comparison of the TPS dose with experimental measurements of the delivered dose[4,5] and, sometimes, with an independent dose calculation algorithm.[6]

In this context, dose comparison tools are needed. Several methods have been developed over time. The dose difference (DD) criterion, the most intuitive of all the methods, proposes to evaluate the difference of dose in each voxel. The distance-to-agreement (DTA) criterion returns the minimum distance between two voxels of the same dose.[7] There are then methods based on both DTA and DD criteria, such as the composite analysis[8,9] or the γ-test.[10] Currently, the γ-test is the most popular method for dose comparison.[11,12] A passing rate is generally computed as the percentage of voxels passing this test. But although the γ-index is widely used in clinical and research environments, it has several well-known drawbacks. First of all, it is not a symmetrical test. Indeed, the results depend on which one of the compared dose distributions is

set as the reference and which one is set as the evaluated dose.[13,14] This results from the fact that, for each reference dose voxel, the evaluated dose is searched to find a voxel having a similar dose to the one of the reference voxel while remaining spatially close enough. Besides the asymmetry issue, the γ-test also has a sensitivity to the resolution of the compared dose maps, especially the evaluated one.[14–17] Indeed, with a lower resolution for the search space of the γ-test, fewer voxels lie within the search radius imposed by the DTA criterion. An extreme example of this situation is when voxel size is larger than the DTA criterion: the γ-test thus boils down to a pixelwise dose difference criterion. Finally, the presence of noise in one or both dose distributions can have an impact as well on the obtained passing rate.[13] Setting the noisy dose as reference leads to an underestimated passing rate. In contrast, choosing the noisy dose as the evaluated one would provide an overestimated passing rate. This effect was theoretically proved for a simplified one-dimensional (1D) case by Graves et al.[18]

The noise dependency of the γ-test may be problematic for QA applications. Indeed, experimental measurements with ionization chambers, radiographic films, or detector arrays generally contain a small noise component. Its impact on the γ passing rate is often limited, but present. To a greater extent, the use of Monte Carlo algorithms as secondary dose verification introduces a more important noise in the compared distribution. Monte Carlo simulations are considered to offer the most accurate dose calculation,[19–21] but their stochastic nature inevitably leads to a statistical noise in the results. Lowering the statistical noise to an acceptable level generally requires one to simulate a huge number of particles, increasing the computation time to an impractical degree for clinical applications. Although several authors reported this γ-index issue[14,17,18] or tried to tackle it,[22] no tractable solution was found yet.

The objective of this paper is to address this issue by analyzing the impact of noise on γ-evaluations and then proposing a method to correct the γ passing rate obtained from noisy distributions. To do so, the γ passing rate is calculated for several quick Monte Carlo simulations computed with various statistical noise levels. The noise-free γ passing rate is then estimated by fitting and extrapolating the data. This allows us to avoid the computation of too costly a Monte Carlo dose while subverting the problem of its noise. The method is illustrated on several proton therapy cases (prostate, lung, brain, liver, and H&N) for the comparison of TPS, Monte Carlo, and measured dose distributions.

2. MATERIALS AND METHODS

2.A. The γ-index method

In the γ-index method, the so-called evaluated dose and reference dose are compared through the use of a criterion made of both DD and DTA. Let D_r denote the reference dose and D_e the evaluated dose. We can therefore define the γ-value at each point $\vec{r}_r$ of the reference dose by

$$\gamma(\vec{r}_r) = \min\{\Gamma(\vec{r}_r, \vec{r}_e)\} \forall \{\vec{r}_e\}, \tag{1}$$

where

$$\Gamma(\vec{r}_r, \vec{r}_e) = \sqrt{\frac{|\vec{r}_e - \vec{r}_r|^2}{\Delta d^2} + \left|\frac{D_e(\vec{r}_e) - D_r(\vec{r}_r)}{D_r(\vec{r}_r)}\right|^2 \frac{100^2}{\Delta D^2}}. \tag{2}$$

$D_r(\vec{r}_r)$ is the value of the reference dose at voxel of coordinates $\vec{r}_r$ and $D_e(\vec{r}_e)$ the value of the evaluated dose at a voxel of coordinates $\vec{r}_e$. The variables Δd and ΔD are the DTA and DD criteria, expressed in mm and %, respectively. They allow the user to fix a tolerance on the search radius and the dose tolerance, respectively. Taking the example of a 3%/3 mm criterion, a necessary condition for a voxel $\vec{r}_r$ to pass the γ-test would be to find a voxel $\vec{r}_e$ such that $\vec{r}_e$ is less than 3 mm away from $\vec{r}_r$ and $D_e(\vec{r}_e)$ has <3% of dose difference with $D_r(\vec{r}_r)$. The value $\Gamma(\vec{r}_r, \vec{r}_e)$ is thus some kind of Euclidean distance between $(\vec{r}_r, D_r(\vec{r}_r))$ and $(\vec{r}_e, D_e(\vec{r}_e))$ whose terms are normalized by a user-chosen tolerance. Therefore, if the minimum value of this variable is less than or equal to 1, the point $\vec{r}_r$ passes the γ-test. This test can thus be seen as the search of a point in the evaluated dose map having a dose similar, to a specified extent, to the reference dose point and being in a given spatial radius around it. At the end, a γ passing rate (GPR) is obtained by calculating the percentage of voxels in the reference dose that have passed the test. A decision can then be made based on a clinical treshold depending on the situation. Typically, a clinical threshold of about 90% success for a 3%/3 mm criterion is required to consider two dose distributions sufficiently similar.[23]

Note that the γ-index, as defined here, is local. A global version of this metric also exists, where the dose difference is normalized by a constant dose such as the maximum reference dose or the prescribed dose.

2.B. Numerical simulations

For the purpose of this study, several proton treatment plans were optimized using the analytical dose calculation algorithm of a commercial TPS, providing noise-free, but approximate dose distributions. In addition, for one of the treatment plans, PSQA measurements were acquired in solid water with a MatriXX PT detector (IBA Dosimetry). Commissioning and validation of the dose calculation algorithm had been previously done for the used beam lines.

In order to obtain noisy dose distributions as well, we used MCsquare,[24] a benchmarked and validated open-source MC code.[21,25] This is a fast algorithm, which allowed us to calculate many dose distributions having various (and sometimes very low) levels of statistical uncertainty. These calculations were based on the CT grid, that is, the MC doses had the same spatial resolution as CT scans, with a voxel size varying between 0.6 and 3 mm. For the PSQA measurement case, a CT scan of the solid water material was used.

Like for the TPS, commissioning and validation of MCsquare had been previously performed for the used beam lines.

In order to quantify the noise level of MC doses, the statistical uncertainty was computed in each voxel using the batch method.[26,27] It was defined voxel-wise as the standard deviation of multiple batch doses divided by the mean dose, which corresponds to the classical relative standard deviation. A general noise level was then obtained, for each plan, by averaging these uncertainties over all pixels having a dose higher than 50% of the maximal dose. The obtained value was then multiplied by 100 in order to get a percentage. In what follows, it will be referred to as $\overline{\sigma}$.

Finally, many γ-tests had to be performed. To this end, we used OpenReggui, an open-source image processing platform for applications in radiotherapy.[28] The γ-index method, in this software, is implemented according to the fast algorithm proposed by Chen et al.[29] All γ-tests in this study were performed over the region receiving at least 10% of the maximum TPS dose.

2.C. Practical considerations in the γ-test

Before using γ-tests on noisy doses, some important limitations should be considered.

2.C.1. Asymmetry

As already stated, the γ-test is an asymmetric tool, meaning that the resulting passing rate will differ depending on which dose distribution is set as the reference. However, there is no physical or mathematical reason to define one dose over another as the reference dose, which complicates the use and the interpretation of the γ-test.

When there is noise in one of the compared dose distributions, the passing rate can be further affected. Monte Carlo algorithms produce noisy results by nature, and it is well known that their statistical uncertainty is approximately proportional to the inverse square root of the simulated particles number.[30]

In order to analyze the combination of both effects mentioned here above, we compared for five different cases analytical doses to Monte Carlo doses with various levels of noise, using γ-tests. These clinical cases were prostate, lung, brain, liver, and H&N tumors. γ-criteria of 2%/2 mm, 3%/3 mm, 4%/4 mm, and 3%/1 mm were used. The MC and TPS doses were set both as reference and then as evaluated doses. These two possible cases will be referred to as MC_{ref} case and MC_{eval} case.

2.C.2. Spatial resolution

It is well known that spatial resolution of compared doses affects the results of a γ-test.[17] In the AAPM Task Group No. 218, the authors recommend to interpolate the evaluated dose so that its resolution is no greater than 1/3 the DTA criterion.[31] This is not an issue in the MC_{ref} case. But when a noisy dose is evaluated, the interpolation just adds meaningless points to the dose map and it affects its overall noise level. In contrast, computation of MC doses with a native resolution as high as 1/3 of the DTA criterion could not always be acceptable in terms of computation time. We believe that, in the scope of our method, the interpolation of MC doses should be maintained, for both pragmatic and common practice reasons. We thus followed AAPM recommendations by resampling evaluated doses using trilinear interpolation.

To remain rigorous, however, we propose a quick analysis on the liver case by comparing GPR for the interpolated case and the MC simulation at high native resolution, in order to evaluate the difference induced in the γ-test in the presence of noise. A resolution of [0.39,0.39,0.5] mm and a γ-criterion of 2%/2 mm are used. We also show the impact of the resolution on the γ-index by observing GPR curves as a function of MC noise for various resolutions.

2.C.3. Normalization of the dose

In this work and, more specifically, in the proposed method, only local γ-tests are performed. However, all presented results could be extrapolated to the global case. The impact of noise on global GPR should nevertheless be studied beforehand. We thus propose here a comparison of the impact of noise on GPR for local and global γ-tests. To this end, we recomputed passing rates obtained with a global γ-index for patients in which we previously observed important noise impact (see Section 2.C.1), that is, lung for the MC_{ref} case and prostate for MC_{eval} case. Doses were normalized to the prescription.

2.D. An adapted use of the γ-index method

In order to reduce the effect of noise on the γ-test, we now propose a method enabling the prediction of the GPR for a low noise MC dose (long computation time) by only computing multiple GPR obtained with various noisy MC doses (short computation time). It consists of a curve fitting followed by an extrapolation: a function linking the GPR to the statistical uncertainty is first derived through an optimization, then this function is evaluated at a low noise level in order to correct the biased GPR.

In practice, seven points $\left(\overline{\sigma}_i,\ \mathrm{GPR}(\overline{\sigma}_i)\right)$, $i = 1,\ldots,7$, are computed, where $\mathrm{GPR}(\overline{\sigma}_i)$ represents the passing rate obtained with a MC dose having a statistical uncertainty $\overline{\sigma}_i$. In clinics, MC simulations are typically performed with a statistical level of 1–2%. Therefore, we chose here to experiment two different cases: either the seven GPR are computed for uncertainties between 1% and 20% or between 2% and 20%. We thus aim at a gain in precision and not in computation time. The smallest statistical uncertainty used (1% or 2%) will be denoted in what follows as $\overline{\sigma}_{\min}$, while the highest one (20%) will be denoted as $\overline{\sigma}_{\max}$. The seven doses needed are extracted at various stages of a single Monte Carlo simulation and do not require additional computation time. Once the γ-tests are performed, a weighted least-squares curve fitting is performed in order to predict the passing rate at a given

uncertainty. This uncertainty must be chosen low enough so that the corresponding GPR can be considered as a reference "denoised" GPR. The value 0.3% was chosen here in order to keep tractable MC computation times during validation. The fitting model is the same for MC_{ref} and MC_{eval} cases and has seven parameters given by the vector $\boldsymbol{c} = [c_1, c_2, c_3, c_4, c_5, c_6, c_7]$. It is given by

$$\begin{aligned} F(\boldsymbol{c}|\overline{\sigma}) = \; & c_1 + c_6\left(c_3\mathrm{erf}\left(\frac{c_3}{\overline{\sigma}}\right) + \frac{\overline{\sigma}}{\sqrt{\pi}}e^{-\left(\frac{c_3}{\overline{\sigma}}\right)^2}\right. \\ & \left. - \; c_2\mathrm{erf}\left(\frac{c_2}{\overline{\sigma}}\right) - \frac{\overline{\sigma}}{\sqrt{\pi}}e^{-\left(\frac{c_2}{\overline{\sigma}}\right)^2}\right) \\ & - c_7\left(c_5\mathrm{erf}\left(\frac{c_5}{\overline{\sigma}}\right) + \frac{\overline{\sigma}}{\sqrt{\pi}}e^{-\left(\frac{c_5}{\overline{\sigma}}\right)^2}\right. \\ & \left. - \; c_4\mathrm{erf}\left(\frac{c_4}{\overline{\sigma}}\right) - \frac{\overline{\sigma}}{\sqrt{\pi}}e^{-\left(\frac{c_4}{\overline{\sigma}}\right)^2}\right) \end{aligned} \tag{3}$$

and was found to be a (parametrized) approximate upper bound of the theoretical GPR evolution as a function of MC noise. The choice of this model is discussed in Appendix A of the supporting information. To perform the fit, the Nelder–Mead direct search algorithm implemented in MATLAB is used. This method was chosen for its efficiency and its calculation speed. But due to the flexibility of our model, which has many parameters, the sensitivity to initial conditions can be quite high in the optimization process. For this reason, we added constraints and made the optimization algorithm explore several initial conditions in order to avoid local minima. The solution for which the objective function is the lowest is then automatically selected. Each optimization being very fast, the entire process takes <10 s.

As the Nelder–Mead method is an unconstrained optimization algorithm, we modified the objective function

$$F_{\mathrm{unconstr}}(\boldsymbol{c}|\overline{\sigma}) = \frac{1}{7}\sum_{j=1}^{7} w_j\left(F(\boldsymbol{c}|\overline{\sigma}_j) - GPR(\overline{\sigma}_j)\right)^2 \tag{4}$$

in order to allow for n constraints $g_i(\boldsymbol{c}|\overline{\sigma}) < 0$, using a penalty function $p(\boldsymbol{c}|\overline{\sigma})$. This way, the new objective function to be optimized is

$$F_{\mathrm{constr}}(\boldsymbol{c}|\overline{\sigma}) = F_{\mathrm{unconstr}}(\boldsymbol{c}|\overline{\sigma}) + p(\boldsymbol{c}|\overline{\sigma}), \tag{5}$$

where the penalty function, weighted by factors α_i, $i = 1,\ldots, n$, is given by

$$p(\boldsymbol{c}|\overline{\sigma}) = \sum_{i=1}^{n} \alpha_i \max\{0, g_i(\boldsymbol{c}|\overline{\sigma})\}^2. \tag{6}$$

Each of the studied cases MC_{ref} and MC_{eval} lead to different constraints. The penalty function p is described hereafter for both cases.

2.D.1. Monte Carlo dose set as the reference dose

Various constraints for the optimization could be selected. In 2013, Graves et al.[18] performed an analysis of the impact of noise on the GPR. Their theoretical conclusion was that in presence of noise in the reference dose, the passing rate was always underestimated. They also numerically validated that result by drawing passing rates as a function of noise for two different patients. Relying on their results, but also on ours (Section 3.A), we chose here to add three different penalties to our objective function, corresponding to three different constraints:

1. *The maximum predicted GPR should be 100%.*
2. *The derivatives should be negative over the extrapolated part of the curve*. This constraint arises from the fact that the GPR decreases when noise increases.
3. *The second derivatives should be negative or close to zero over the extrapolated part of the curve*. This constraint is derived from our results and observations of Graves et al.'s results.

Taking these constraints into account, the penalty term is finally chosen as

$$\begin{aligned} p(\boldsymbol{c}|\overline{\sigma}) = \frac{1}{l}\Bigg(& 100\max\{0, F(\boldsymbol{c}|0.05 \le \overline{\sigma} \le \overline{\sigma}_{\min}+1) - 100\}^2 \\ & + 40\max\left\{0, \frac{\partial F}{\partial \overline{\sigma}}(\boldsymbol{c}|0.05 \le \overline{\sigma} \le \overline{\sigma}_{\min}+1)\right\}^2 \\ & + 20\max\left\{0, \frac{\partial^2 F}{\partial \overline{\sigma}^2}(\boldsymbol{c}|0.05 \le \overline{\sigma} \le \overline{\sigma}_{\min}+1) - 0.5\right\}^2 \\ & + 20\max\left\{0, -\frac{\partial^2 F}{\partial \overline{\sigma}^2}(\boldsymbol{c}|0.05 \le \overline{\sigma} \le \overline{\sigma}_{\min}+1) - 4\right\}^2\Bigg), \end{aligned} \tag{7}$$

where l is the number of points in which constraints are evaluated and the α_i in Eq. (6) are chosen to emphasize the most important penalties.

2.D.2. Monte Carlo dose set as evaluated dose

With similar reasoning as for the MC_{ref} case, the following constraints are applied during the optimization process of the MC_{eval} case:

1. *The maximum predicted GPR should be 100%.*
2. *The derivatives should be positive over the extrapolated part of the curve.*
3. *The second derivatives should be negative or close to zero over the extrapolated part of the curve.*

These constraints lead us to a final penalty term equal to the one given in Eq. (7), except for the sign of the first derivatives that need to be positive and not negative, resulting in a change of sign in the second constraint.

2.E. Application and evaluation of the method

Our method is tested on nine proton therapy plans. The five first ones are the ones already mentioned in Section 2.C. To these are added other brain, lung, and H&N cases, along

with the previously mentioned PSQA case (lung). For this last plan, one γ-test between the measurement and MC dose was performed for each of its two beams (left posterior oblique (LPO) and posterior (POST)). As measurements only provide 2D information, the MC_{eval} case leads to 2D/3D γ-tests since the search can be done in 3D around the slice corresponding to the measurement. In contrast, for the MC_{ref} case, γ-tests are 2D/2D since it would make no sense to use a whole 3D MC volume as a reference when all we have is a 2D evaluated dose, corresponding to a single slice of the reference dose. In the latter case, the slice of the MC dose is selected so that it corresponds to the theoretical measurement depth. For the other plans, 3D TPS and MC doses are compared. γ-criteria of 2%/2 mm, 3%/3 mm, 4%/4 mm and 3%/1 mm are used.

Having at our disposal GPR corresponding to MC statistical noises of 1%, 2%, 4%, 7%, 8%, 13%, and 20% ($\overline{\sigma}_{min} = 1\%$) or 2%, 3%, 4%, 7%, 8%, 13%, and 20% ($\overline{\sigma}_{min} = 2\%$), the curve fittings were done by giving more weight to low noise passing rates (w = [40,30,20,10,5,1,1]).

As explained in Section 2.D, to avoid local minima, the optimization algorithm explores multiple solutions by randomly sampling initial conditions several times. Due to this random sampling, the optimization results may vary from one run to another. In order to evaluate this statistical deviation, we applied our method 60 times for each case and report here the average predicted passing rates. Along with their standard deviation, the GPR mean absolute error with respect to the 0.3% uncertainty dose is also given, as well as the gain obtained with our method which was defined as

$$\text{Gain} = \left|\text{GPR}_{\text{true},0.3\%} - \text{GPR}_{\text{true},\overline{\sigma}_{\text{min}}}\right| - \left|\text{GPR}_{\text{true},0.3\%} - \text{GPR}_{\text{predicted},0.3\%}\right|, \quad (8)$$

where $\text{GPR}_{\text{true},0.3\%}$ is the calculated GPR for 0.3% of statistical uncertainty, $\text{GPR}_{\text{true},\overline{\sigma}_{\text{min}}}$ is the calculated passing rate obtained for the fitting point having the smallest uncertainty among all, and $\text{GPR}_{\text{predicted},0.3\%}$ is the passing rate predicted by our method for the reference uncertainty 0.3%. This gain therefore represents the difference between the errors on the passing rate obtained with the classical γ-index (with a MC uncertainty $\overline{\sigma}_{min}$) and with our method, if we consider the GPR for a noise level of 0.3% as being the true GPR. A positive gain therefore corresponds to an improvement by our method compared to the classical γ-index where the MC dose is calculated for a noise level $\overline{\sigma}_{min}$, while a negative gain would indicate a worsened situation.

3. RESULTS

3.A. Practical considerations in the γ-test

3.A.1. Asymmetry and noise impact

The asymmetry issue of the γ-test is illustrated on Fig. 1 for all five patients, with a γ-criterion of 2%/2 mm. The figure shows the passing rate curves obtained as a function of the noise level in the MC dose for both possible cases: MC dose defined as the reference or as the evaluated dose. Brain and H&N cases, particularly, show that MC_{eval} and MC_{ref} curves do not tend to a same GPR value as the uncertainty tends to zero.

Figure 2 summarizes the error made on the GPR for all γ-criteria and five patients, according to the status of the MC dose and its level of noise. This error is defined as the difference between the GPR for a given $\overline{\sigma}$ ($\text{GPR}_{\overline{\sigma}}$) and the GPR for the lowest available uncertainty ($\text{GPR}_{\overline{\sigma}_0}$). A more detailed table with all results is also shown in Appendix B (Table S1) of the supporting information.

For the case where the Monte Carlo dose was defined as the reference dose in the γ-test, the effect of GPR underestimation due to noise presence is clearly observed; the GPR curves in Fig. 1 are increasing when the statistical uncertainty decreases and are very similar, irrespective of the patient and the criterion. However, the magnitude of the underestimation varies a lot. For a noise level of 1%, GPR underestimation can be negligible (below 1%) or very important (going up to 8%), depending on the patient and the criterion. For a noise level of 2%, it can even go up to almost 20%. Generally, the smaller the DTA criterion, the stronger the impact of noise, which makes sense since the γ-index metric consists of a DD criterion with a DTA relaxation.

For the case where the Monte Carlo distribution is the evaluated dose in the γ-test, the effect of GPR overestimation due to noise is clear as well. However, the effect is limited when compared doses are very similar, since the maximum possible GPR is 100%. This is why, for some cases, the impact of noise is very small. The maximum GPR overestimation over all five patients and criteria reaches 5.27% for $\overline{\sigma} = 1\%$ and 11.38% for $\overline{\sigma} = 2\%$.

A more detailed analysis of the noise impact is given in Appendix B of the supporting information. We discuss there what might affect the magnitude of GPR error due to noise.

3.A.2. Spatial resolution

Figure 3 shows GPR curves for various resolutions of the evaluated dose in the liver case. The DTA criterion being here 1 mm, only the highest resolution considered ([0.29, 0.29, 0.33] mm) fits in the criteria of AAPM TG-218.[31] According to Fig. 3, it seems justified to require such a high resolution, especially in presence of noise. Notice that with the lowest resolution used here, the DTA has no impact at all since it is smaller than the voxel size in each direction. If we do not consider this curve, low noise GPR are actually quite similar for all higher resolutions, that is, the impact of resolution is still present but minimal.

Figure 4 shows GPR evolutions for the liver MC_{eval} case, comparing two different γ-index computations: either the evaluated (MC) dose is interpolated to a resolution of [0.39, 0.39, 0.5] mm or it is directly computed with this resolution. For each considered γ-criterion, we observe a similar behavior: the interpolation results in a slightly lower GPR, regardless of the noise level.

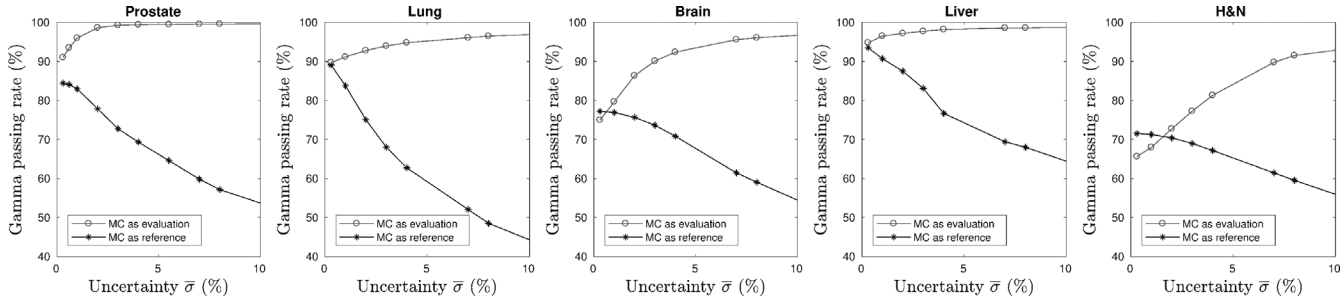

FIG. 1. Gamma passing rate for 2%/2 mm as a function of mean statistical uncertainty, for prostate, lung, brain, liver, and H&N cases (from left to right). The darker curve shows the case where MC dose is set as reference while the lighter one shows the case where TPS dose is set as reference.

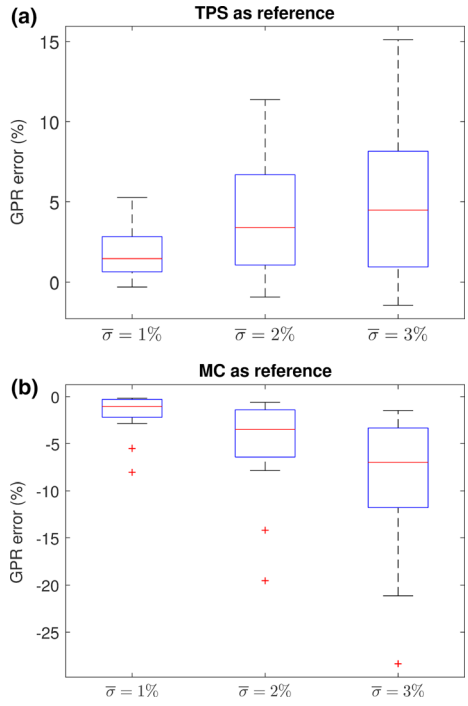

FIG. 2. GPR errors summarized for five patients (prostate, lung, brain, liver, and H&N) and four γ-criteria (2%/2, 3%/3, 4%/4, and 3%/1 mm), for (a) MC_{eval} case and (b) MC_{ref} case.

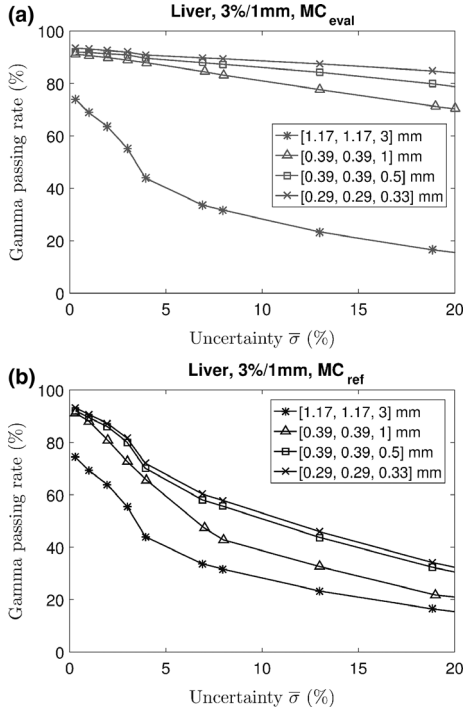

FIG. 3. Comparison of several spatial resolutions of the evaluated dose for 3%/1 mm GPR as a function of mean statistical uncertainty in the liver case, for (a) MC_{eval} case and (b) MC_{ref} case.

3.A.3. Normalization of the dose

Figure 5 shows the GPR evolution as a function of MC noise for various cases: the top row is the MC_{ref} lung case, comparing global and local γ-tests; the bottom row shows the same but for the MC_{eval} case with the prostate patient.

The choice of patients for this analysis was based on the importance of the GPR error observed for given noise levels. We indeed observed in Section 3.A.1 that our lung patient shows an important underestimation of the local GPR in the MC_{ref} case while the prostate patient leads to high GPR overestimation in the MC_{eval} case. We selected this way the worst cases. As we could expect, since the dose difference is normalized by a high constant dose in the γ-metric, we observe in Fig. 5 that a global γ-index metric always leads to a better GPR. Moreover, it lessens the impact of noise. For the MC_{eval} case, there is almost no GPR difference between various low noise levels such as 0.3%, 1% or 2%. For the MC_{ref} case, the GPR underestimation for noise levels of 1% or 2% can reach 4.1%. This value remains nevertheless way below GPR errors observed for local γ-metrics.

3.B. An adapted use of the γ-index method

3.B.1. Monte Carlo dose set as reference dose

Figure 6 illustrates the method by showing a resulting fit along with the absolute value of the error in each point for the case Lung 1 and each of the four γ-criteria considered. The figure is zoomed over a range of uncertainties between 0.3% and 20%, in semi-logarithmic scale.

Figure 7 summarizes the results obtained over all nine patients (10 γ-tests with the PSQA measurements case) and

γ-criteria when $\overline{\sigma}_{\min} = 1\%$. Figure 7(a) compares the GPR mean errors before and after applying the method while Fig. 7(b) shows the resulting mean gains. We see that our method allowed us to strongly reduce GPR errors in such a way that most of them are below 1%. Moreover, the maximum error goes down from 8.03% to 1.96%. Gains are mostly between 0% and 2%. However, we understand from the few high gains that the method successfully corrected the highest GPR errors. Some negative gains are also observed, going down to -0.87%.

Similarly to Fig. 7, Fig. 8 summarizes the results obtained over all patients and γ-criteria when $\overline{\sigma}_{\min} = 2\%$. We see this time that new GPR errors are mostly between 0% and 3%. The maximum error is reduced from 19.53% to 6.37%. It remains an important error, however it corresponds to a gain of 12.96%. Other gains are mostly between 0% and 7%.

More detailed results are given in Appendix C of the supporting information. They provide GPR for 0.3% and $\overline{\sigma}_{\min}$, mean GPR obtained, mean errors, mean gains, and corresponding standard deviations per patient and per γ-criterion, including for the PSQA measurement case.

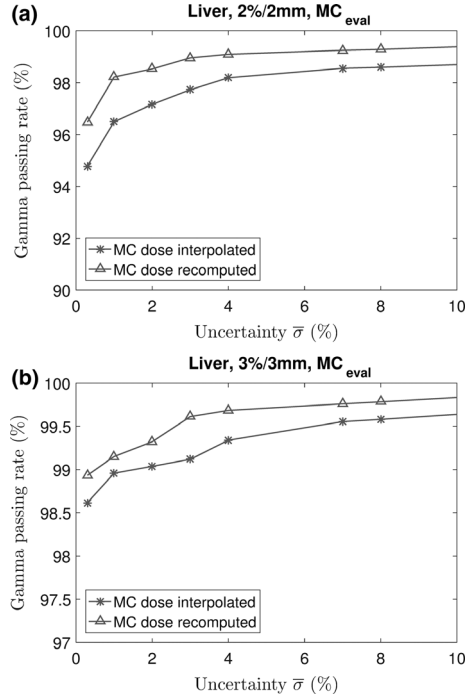

FIG. 4. Comparison of MC evaluated doses interpolated and directly computed at desired resolution in the liver case, for (a) 2%/2 mm GPR and (b) 3%/3 mm GPR. Resolution of the MC dose is (0.39, 0.39, 0.5) mm.

3.B.2. Monte Carlo dose set as the evaluated dose

Figure 9 shows a resulting fit along with the absolute error in each point for the case H&N 1 and each of the four considered γ-criteria, in semi-logarithmic scale. We can see in this case that the fitting worked quite well.

Figure 10(a) shows the GPR errors before and after applying our method, for $\overline{\sigma}_{\min} = 1\%$. All nine patients and γ-criteria are included. Figure 10(b) gives the corresponding gains obtained with the proposed method. We can observe that the repartition of errors after applying our method is concentrated around zero. Most of them are below 1% and the maximum one is 1.89%. However, the initial impact of noise was less important than in the MC_{ref} case, which leads to gains no higher than 3%.

Similarly to Fig. 10, Fig. 11 summarizes the results obtained over all nine patients and γ-criteria when

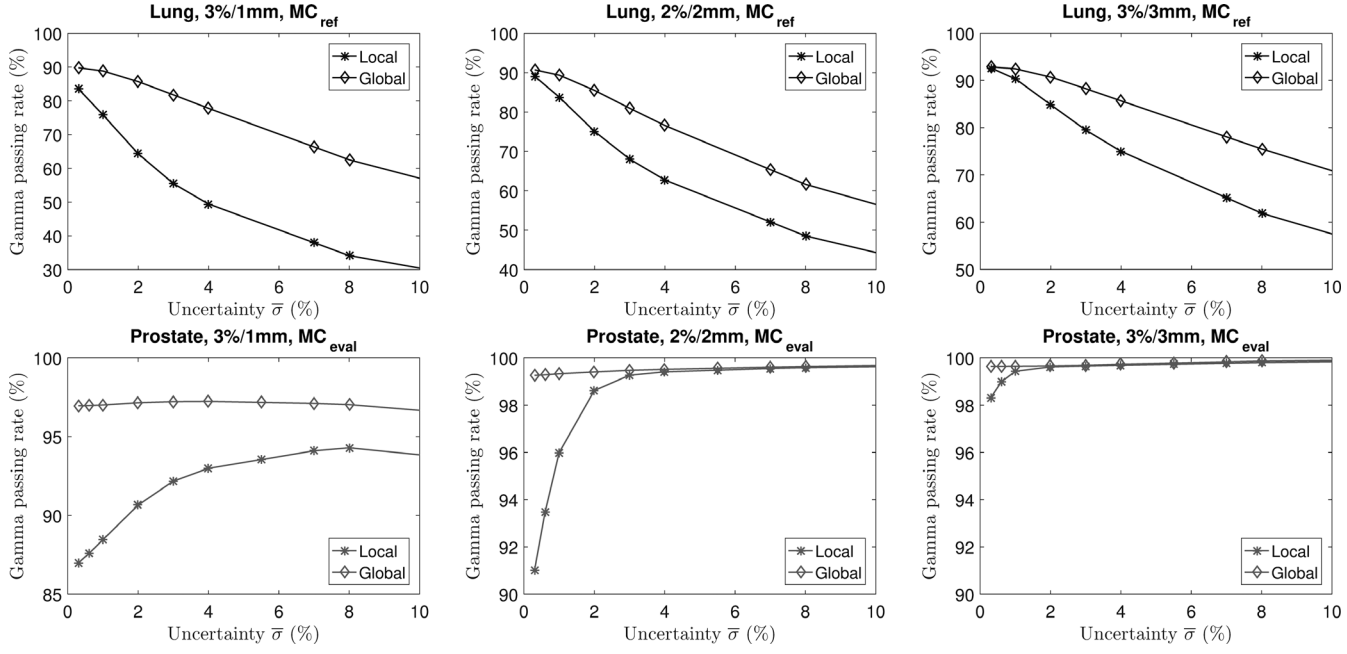

FIG. 5. Comparison of global and local γ-tests: GPR as a function of mean statistical uncertainty, for lung (above) when MC dose is set as reference and prostate (below) when MC dose is evaluated (3%/1 mm, 2%/2 mm, and 3%/3 mm from left to right).

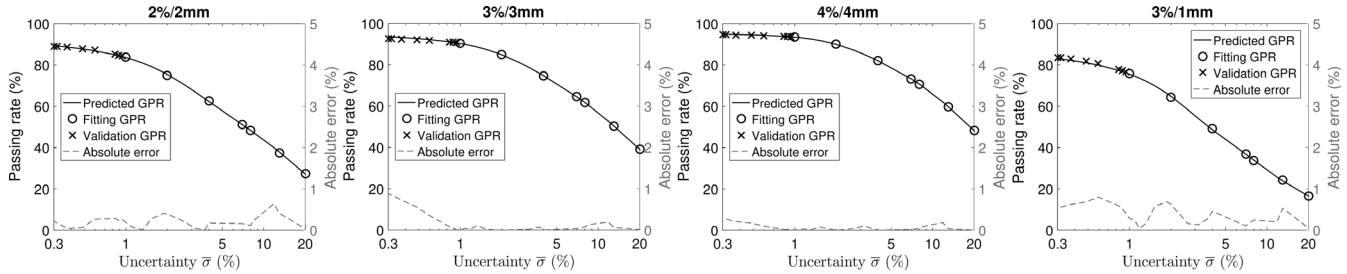

FIG. 6. MC_{ref} case : resulting fitting and corresponding absolute error for the case Lung 1 with γ-criterion 2%/2, 3%/3, 4%/4, and 3%/1 mm, when $\overline{\sigma}_{min} = 1\%$. The circles show the points used for fitting, the crosses show the true passing rates used to validate the prediction, the plain curve shows the predicted passing rate and the dashed line the absolute error.

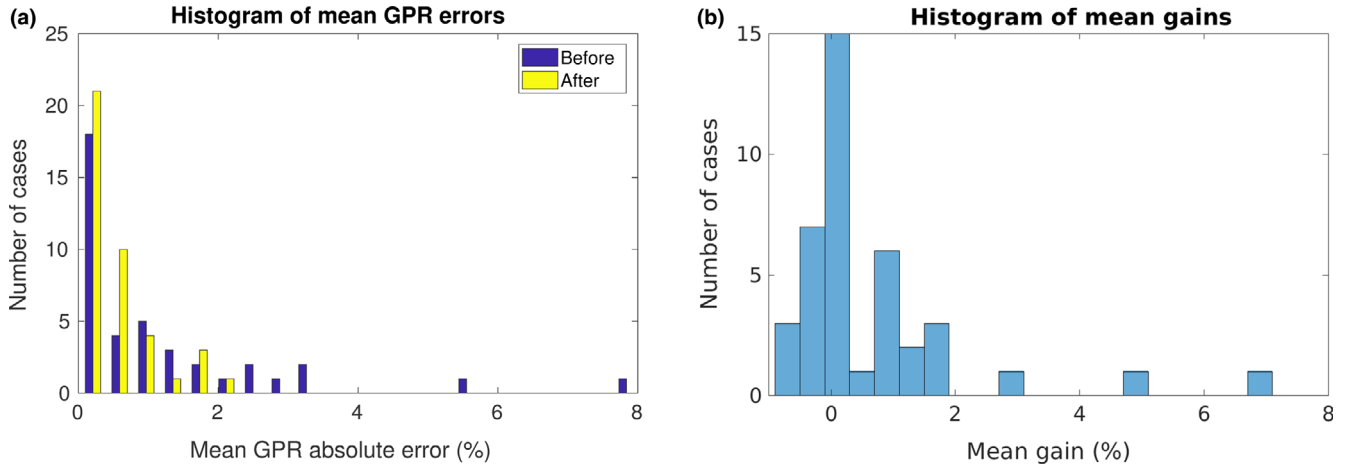

FIG. 7. Summary of the results using the adapted γ-index in the MC_{ref} case with $\overline{\boldsymbol{\sigma}}_{\mathbf{min}} = \mathbf{1}\%$, for all nine patients and γ-criteria: (a) repartition of mean absolute errors, computed before and after applying the method and (b) repartition of mean gains.

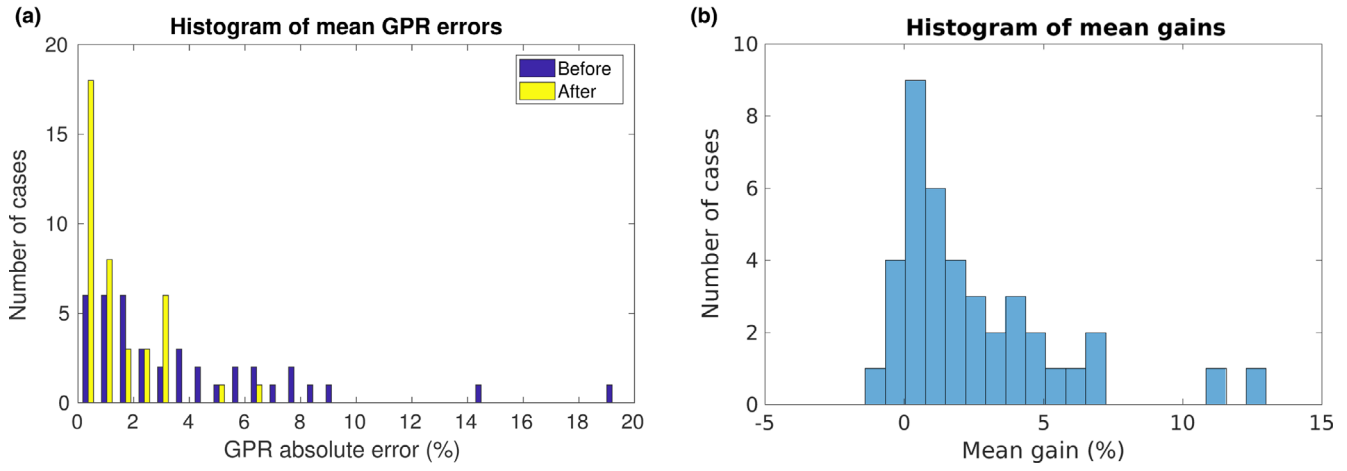

FIG. 8. Summary of the results using the adapted γ-index in the MC_{ref} case with $\overline{\boldsymbol{\sigma}}_{\mathbf{min}} = \mathbf{2}\%$, for all nine patients and γ-criteria: (a) repartition of mean absolute errors, computed before and after applying the method and (b) repartition of mean gains.

$\overline{\sigma}_{min} = 2\%$. We can see here again that our method allows us to reduce GPR errors down to 2% in most cases, with a maximum error of 6.24% instead of 11.38%. Most gain are between 0% and 3% but can go up to almost 10%.

Similarly to the MC_{ref} case, more detailed results are given in Appendix C of the supporting information.

4. DISCUSSION

As already mentioned in several papers,[13,14,17,18] the noise in dose distributions can heavily impact the results of a dose comparison performed with the γ-index method. As shown in Section 3.A, noise in the reference dose results in an

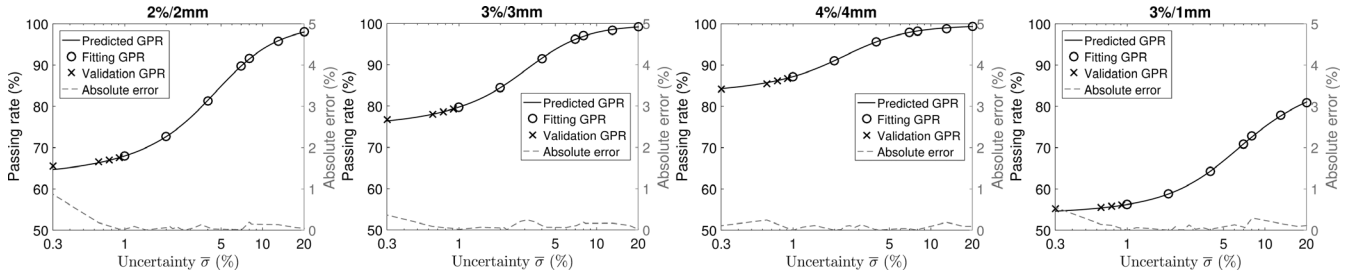

FIG. 9. MC_{eval} case : resulting fitting and corresponding absolute error for the H&N case with γ-criterion 2%/2, 3%/3, 4%/4, and 3%/1 mm, when $\overline{\sigma}_{min} = 1\%$. The circles show the points used for fitting, the crosses show the true passing rates used to validate the prediction, the plain curve shows the predicted passing rate, and the dashed line the absolute error.

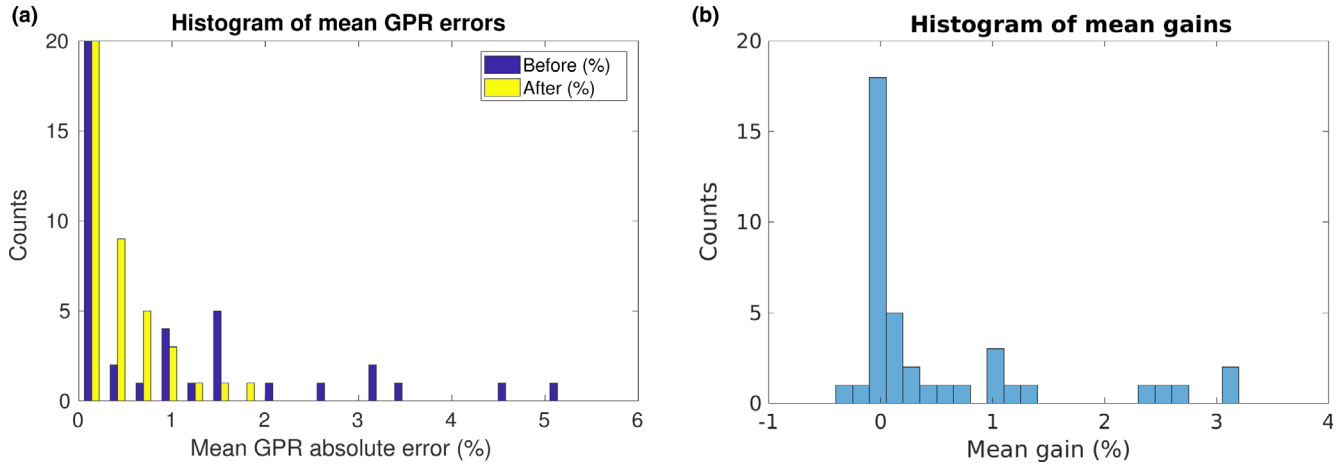

FIG. 10. Summary of the results using the adapted γ-index in the MC_{eval} case with $\overline{\sigma}_{min} = 1\%$, for all 9 patients and γ-criteria: (a) repartition of mean absolute errors, computed before and after applying the method and (b) repartition of mean gains.

underestimated passing rate. In contrast, noise in the evaluated dose typically produces an overestimated passing rate. In our study, for a classical 3%/3 mm criterion, the GPR underestimation between 1% and the smallest available uncertainty for the MC_{ref} case could go up to 8.03%. Moreover, a smaller DTA criterion usually led to a higher error, which makes sense since the γ-index metric consists of a DD criterion with a DTA relaxation. For the MC_{eval} case, the GPR error went up to 5.27%.

These errors being nevertheless very patient-specific, it is difficult to evaluate the accuracy of a given γ-test. We did not find here (see Appendix B of the supporting information) any clear criterion to determine the severity of noise impact. Although there might be a slight dependence on the true passing rate, this value is normally not known in practice. Note that this trend was also observed by Huang et al.[17] for a global γ-index. A feature not studied in this paper is the proportion of high dose gradients in the compared doses. Low and Dempsey showed in their paper of 2003 that high gradients regions in the doses led to less impact of noise on the γ-metric.[13] However, this does not provide any straightforward way to quantify noise effect on the GPR.

A more practical way to determine if a γ-evaluation is impacted by noise could be to perform γ-tests for two different (and close) noise levels, such as 1% and 2% or 2% and 3%, to then compute the difference between obtained GPR. This could already give a good idea of the magnitude of the GPR error.

Another factor that can impact the γ passing rate and was discussed here is the spatial resolution of the evaluated dose distribution. Indeed, we should in theory have a continuous search space when performing a γ-test. But due to the discrete nature of the dose image, this is not the case and the evaluated dose thus needs to be interpolated. The current recommendation, from AAPM TG-218,[31] is to always keep its resolution below 1/3 of the DTA criterion. In this paper, we compared GPR as a function of the noise level for various resolutions and observed that a coarser resolution worsened the impact of noise. The effect of interpolating a noisy (evaluated) dose was also studied. It was found to generate slightly lower GPR for γ-criteria such as 2%/2 and 3%/3 mm. This stems from the fact that the interpolation process smooths the dose as a side effect and, therefore, the estimated noise level is too high. The observed effect is thus mainly the GPR overestimation in presence of noise in the evaluated dose, although part of it might also be due to the interpolation itself, which generates artificial dose values. When decreasing the DTA criterion down to 1 mm (see Appendix B of the supporting information), however, a different effect was observed. It resulted in a decreasing GPR instead of an

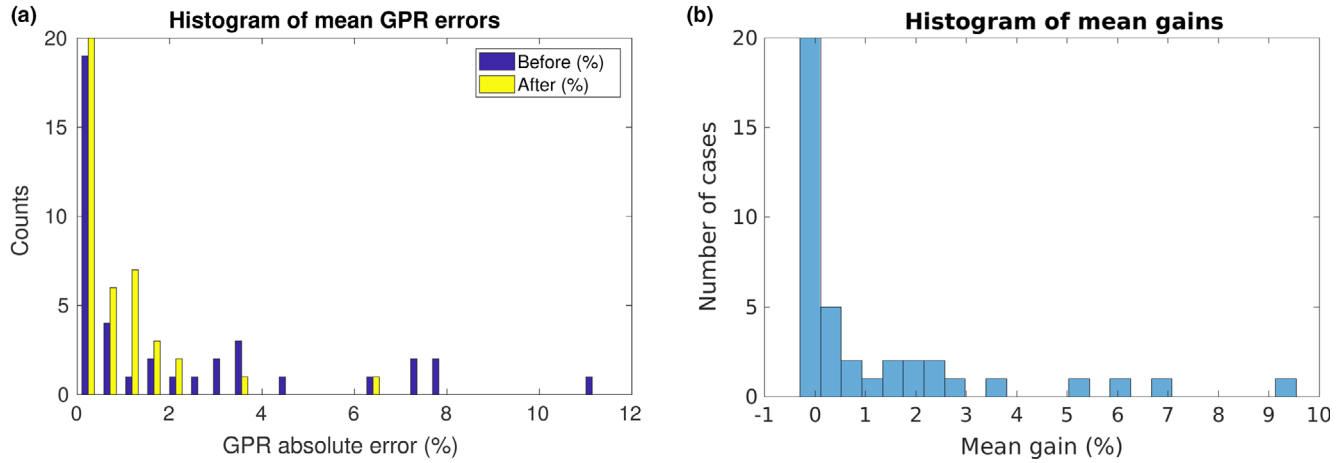

FIG. 11. Summary of the results using the adapted γ-index in the MC_{eval} case with $\overline{\sigma}_{min} = 2\%$, for all 9 patients and γ-criteria: (a) repartition of mean absolute errors, computed before and after applying the method and (b) repartition of mean gains.

increasing one as when the MC dose is computed with a high native resolution. Although only one patient was considered here, we believe that resampling noisy doses should be avoided when possible.

A last feature that needs to be considered to enable a practical use of the γ-metric in presence of noise is the dose normalization, that is, the global/local nature of the γ-test. It was found here that the impact of noise is least when normalization is performed globally, as expected, since the dose difference is then normalized by a high dose constant in the γ-metric. For a MC dose having a noise level of 1% or 2%, noise effect becomes essentially negligible.

In order to overcome these issues of the γ-test for the comparison of MC doses and noise-free doses, we introduced a correction which allowed us to take noise in consideration. The proposed method is an extrapolation based on the computation of several γ-tests of multiple MC doses having various high levels of statistical uncertainty, the lowest being denoted as $\overline{\sigma}_{min}$. A more precise GPR can this way be obtained by extrapolation without increasing the MC computation time. The method was considered for both MC_{ref} and MC_{eval} cases.

When we set the MC dose as reference in the γ-tests for a $\overline{\sigma}_{min}$ of 1%, good results are achieved by our method. Mean GPR errors, initially between 0% and 8%, are reduced down to maximum 2%, leading to gains up to 7%. Standard deviations are very low (see Appendix C of the supporting information). However, the method does not systematically allow for an improvement because the GPR, at 1% of uncertainty, is already sometimes very close to the true GPR. This causes some losses in accuracy, which remain nevertheless weak in magnitude and correspond to small GPR errors. When applying the adapted γ-index with a $\overline{\sigma}_{min}$ equal to 2%, mean GPR errors go down from maximum 19% (for a regular γ-test) to maximum 6%, most of them being actually between 0% and 3%. Thus, using the method for a higher $\overline{\sigma}_{min}$ still reduces strongly GPR errors but does not guarantee such a good GPR accuracy.

For the MC_{eval} case with a $\overline{\sigma}_{min}$ of 1%, broadly speaking, mean gains are lower because initial GPR errors are lower. They go up to 3%, reducing mean GPR errors from maximum 5% down to <2%. However, mean errors are mainly between 0% and 1%. For the same reason as previously mentioned, a few negative gains of small amplitude are observed. For a $\overline{\sigma}_{min}$ of 2%, results similar to the MC_{ref} case can be observed, except for the standard deviations (see Appendix C of the supporting information), which are sometimes higher but then correspond to high gains. It seems that, in both cases, using the method for a higher $\overline{\sigma}_{min}$ still reduces strongly GPR errors but does not guarantee such a good GPR accuracy.

Our results can also be interpreted in terms of computation time. To this end, let us imagine that a noise level as low as 0.3% is required to consider obtained time gain. For the 10 cases considered, the use of the adapted γ-index enabled to reduce the Monte Carlo computation time by an approximate factor between 4 and 34 for a $\overline{\sigma}_{min}$ of 1%, depending on the patient. This means our method was at least three times faster than the usual γ-index. For a $\overline{\sigma}_{min}$ of 2%, the factor of gain in computation time reached values between 5.2 and 157. For example, the MC simulation for the case Lung 1 required 80 min to reach an uncertainty as small as 0.3%, but only 7 and 1.8 min, respectively, to attain 1% and 2% of noise. The duration of a single γ-test remains below 1 min, depending on the patient and the resolution of the reference dose.

Based on this first part of the discussion, some recommendations can now be made on how to perform appropriate γ-tests and interpret them correctly when one of the compared dose maps comes from a Monte Carlo calculation.

We first would like to emphasize how important it is to mention how exactly the γ-test is done when reporting results. Information such as the spatial resolution of compared doses, noise level of the MC dose, normalization factor, and status of each dose (reference or evaluated) should always be mentioned when reporting γ-evaluations in the literature.

Then, we believe that when comparing a MC dose to an analytical one, the MC dose should preferably be set as reference. Several reasons support this statement. First, the GPR is in that case always underestimated, while for the MC_{eval} case it is usually overestimated. It is thus safer, especially in a clinical context, to prefer the MC_{ref} option. This is also a good choice to reduce the need of a noise correction method: if the computed GPR is already high enough, no need to worry about its true value. Secondly, the spatial resolution of the evaluated dose strongly affects the GPR and interpolation of the dose distribution is therefore needed. However, applying upsampling to a noisy dose can also lead to biased passing rates, as shown in Section 3.A and Appendix B of the supporting information. Finally, the MC tool is supposed to provide doses more accurately than analytical algorithms, and it thus makes sense to set MC doses as the references.

When comparing MC doses to measurements, however, our last argument does not hold anymore. Moreover, when setting the MC dose as reference to compare it to a 2D measurement, the search performed by the γ-test becomes 2D and not 3D anymore, as explained in Section 2.E. A way around this could be to use the MC dose as evaluated distribution while adapting the GPR passing treshold or using a correction method such as the one proposed here.

Another recommendation would be to impose a lower noise level when computing a local γ-index, which is more sensitive to noise than a global one. Based on our results, an uncertainty of 1% seems quite reasonable.

Finally, for the spatial resolution of the evaluated dose, recommendations from AAPM TG-218 should be followed,[31] that is, the voxel size in the evaluated dose should remain below 1/3 of the DTA criterion.

This being stated, the general framework of the proposed method could, in our opinion, lead to a good compromise between accuracy and computation time for a QA process involving a γ-index evaluation based on Monte Carlo dose distributions. However, there might still be some work to be done in order to enable an efficient use of the method. For instance, in order to avoid negative gains occurring when no noise correction is needed, we could imagine to compute beforehand two GPR for two different levels of noise (such as 1% and 2% or 2% and 3%) and to deduce from their difference if the adapted γ-index should be used or not. Likewise, it could be good to explore a way to reduce the variability observed when $\overline{\sigma}_{min} = 2\%$ (in particular for the MC_{eval} case); an option might be to look for another fitting model having fewer parameters or approximating better the shape of GPR curves as a function of noise. This could be the scope of further investigation. Another limitation of our method is the case in which we would compare two noisy doses; this is currently not possible since there is no certitude about the impact of noise produced on GPR. A full study should hence be performed to first investigate noise impact then adapt the correction method if needed. Finally, the proposed method was only tested on PBS proton plans here and it might be good to verify if it still applies to other treatment modalities such as conventional radiotherapy. Nonetheless, we believe it should be the case as the shape of the passing rate evolution curve with the statistical MC uncertainty should be independent of the radiation type.

5. CONCLUSIONS

The dose comparison workflow should be adapted when it involves Monte Carlo doses. Noise impacts γ passing rates and worsens the impact of other factors affecting the γ-index method.

We proposed here an adapted use of the usual γ-index method for Monte Carlo dose comparison. The idea is to predict the γ passing rate for an almost noiseless MC dose based on several passing rates computed for MC doses having various high levels of noise, with an aim at accuracy gain. This allowed us to reduce the Monte Carlo computation time that would be required to reach a noise-free GPR.

ACKNOWLEDGMENTS

Marie Cohilis is supported by the Télévie Grant from the Belgian "Fonds National pour la Recherche Scientifique" F.R.S-FNRS (Grant No. 7450517F). Kevin Souris is supported by a research grant from Ion Beam Application (IBA s.a., Louvain-la-Neuve, Belgium). John A. Lee is a Research Associate with the Belgian fund of scientific research (F.R.S.-FNRS).

CONFLICT OF INTEREST

There is no conflict of interest declared in this article.

[a)]Author to whom correspondence should be addressed. Electronic mail: marie.cohilis@uclouvain.be.

REFERENCES

1. Clark CH, Bidmead AM, Mubata CD, Harrington KJ, Nutting CM. Intensity-modulated radiotherapy improves target coverage, spinal cord sparing and allows dose escalation in patients with locally advanced cancer of the larynx. *Radiother Oncol*. 2004;70:189–198.
2. Zhang X, Li Y, Pan X, et al. Intensity-modulated proton therapy reduces the dose to normal tissue compared with intensity-modulated radiation therapy or passive scattering proton therapy and enables individualized radical radiotherapy for extensive stage IIIB non-small-cell lung cancer: a virtual clinical study. *Int J Radiat Oncol Biol Phys*. 2010;77:357–366.
3. Ezzell G, Galvin J, Low D, et al. Guidance document on delivery, treatment planning, and clinical implementation of IMRT: report of the IMRT subcommittee of the AAPM radiation therapy committee. *Med Phys*. 2003;30:2089–2115.
4. Low DA, Mutic S, Dempsey JF, et al. Quantitative dosimetric verification of an IMRT planning and delivery system. *Radiother Oncol*. 1998;49:305–316.
5. Van Esch A, Bohsung J, Sorvari P et al. Acceptance tests and quality control (QC) procedures for the clinical implementation of intensity modulated radiotherapy (IMRT) using inverse planning and the sliding window technique: experience from five radiotherapy departments. *Radiother Oncol*. 2002;65:53–70.
6. Meier G, Besson R, Nanz A, Safai S, Lomax A. Independent dose calculations for commissioning, quality assurance and dose reconstruction of PBS proton therapy. *Phys Med Biol*. 2015;60:2819–2836.

7. Van Dyk J, Barnett RB, Cygler JE, Shragge PC. Commissioning and quality assurance of treatment planning computers. *Int J Radiat Oncol Biol Phys*. 1993;26:261–273.
8. Shiu AS, Tung S, Hogström KR, et al. Verification data for electron beam dose algorithms. *Med Phys*. 1992;19:623–636.
9. Harms WB, Low DA, Wong JW, Purdy JA. A software tool for the quantitative evaluation of 3D dose calculation algorithms. *Med Phys*. 1998;25:1830–1836.
10. Low DA, Harms WB, Mutic S, Purdy JA. A technique for the quantitative evaluation of dose distributions. *Med Phys*. 1998;25:656–661.
11. Agazaryan N, Solberg TD, DeMarco JJ. Patient specific quality assurance for the delivery of intensity modulated radiotherapy. *J Appl Clin Med Phys*. 2003;4:40–50.
12. Wendling M, Zijp LJ, McDermott LN, et al. A fast algorithm for gamma evaluation in 3D. *Med Phys*. 2007;34:1647–1654.
13. Low DA, Dempsey JF. Evaluation of the gamma dose distribution comparison method. *Med Phys*. 2003;30:2455–2464.
14. Low DA, Gamma dose distribution evaluation tool. *J Phys Conf Ser*. 2010;250:349–359.
15. Depuydt T, Van Esch A, Huyskens DP. A quantitative evaluation of IMRT dose distributions: refinement and clinical assessment of the gamma evaluation. *Radiother Oncol*. 2002;62:309–319.
16. Jiang SB, Sharp GC, Neicu T, Berbeco RI, Flampouri S, Bortfeld T. On dose distribution comparison. *Phys Med Biol*. 2006;51:759–776.
17. Huang JY, Pulliam KB, McKenzie EM, Followill DS, Kry SF. Effects of spatial resolution and noise on gamma analysis for IMRT QA. *J Appl Clin Med Phys*. 2014;15:93–104.
18. Graves YJ, Jia X, Jiang SB. Effect of statistical fluctuation in Monte Carlo based photon beam dose calculation on gamma index evaluation. *Phys Med Biol*. 2013;58:1839–1853.
19. Jelen U, Alber M. A finite size pencil beam algorithm for IMRT dose optimization: density corrections. *Phys Med Biol*. 2007;52:617–633.
20. Hissoiny S, Ozell B, Bouchard H, Després P. GPUMCD: a new GPU-oriented Monte Carlo dose calculation platform. *Med Phys*. 2011;38:754–764.
21. Huang S, Souris K, Li S, et al. Validation and application of a fast Monte Carlo algorithm for assessing the clinical impact of approximations in analytical dose calculations for pencil beam scanning proton therapy. *Med Phys*. 2018;45:5631–5642.
22. Dhakal TR, Yepes P. A symmetric probabilistic gamma-index for Monte Carlo dose comparisons. *Phys Med Biol*. 2014;59:N153–N161.
23. Ezzell GA, Burmeister JW, Dogan N, et al. IMRT commissioning: multiple institution planning and dosimetry comparisons, a report from AAPM Task Group 119. *Med Phys*. 2009;36:5359–5373.
24. Souris K, Lee JA, Sterpin E. Fast multipurpose Monte Carlo simulation for proton therapy using multi- and many-core CPU architectures. *Med Phys*. 2016;43:1700–1712.
25. Huang S, Kang M, Souris K, et al. Validation and clinical implementation of an accurate Monte Carlo code for pencil beam scanning proton therapy. *J Appl Clin Med Phys*. 2018;19:558–572.
26. Andreo P. Monte Carlo techniques in medical radiation physics. *Phys Med Biol*. 1991;36:861–920.
27. Seco J, Fragoso M. Patient Dose Calculation; 2013:111–124.
28. OpenReggui, https://openreggui.org/.
29. Chen M, Lu W, Chen Q, Ruchala K, Olivera G. Efficient gamma index calculation using fast Euclidean distance transform. *Phys Med Biol*. 2009;54:2037–2047.
30. Fippel M. Basics of Monte Carlo simulations; 2013:17–28.
31. Miften M, Olch A, Mihailidis D, et al. Tolerance limits and methodologies for IMRT measurement-based verification QA: recommendations of AAPM Task Group No. 218. *Med Phys*. 2018;45:e53–e83.
32. Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilitá. *Pubblicazioni del R. Istituto superiore di scienze economiche e commerciali di Firenze*. 1936;8:3–62.
33. Ng EW, Geller M. A table of integrals of the error functions. *J Res Natl Bur Stand Sec B*. 1969;73B:1–20.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix S1: Supplementary Material.