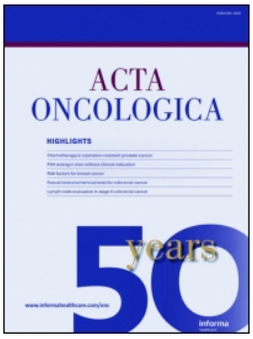


Acta Oncologica

ISSN: 0284-186X (Print) 1651-226X (Online) Journal homepage: https://www.tandfonline.com/loi/ionc20

# Intensity modulated proton therapy plan generation in under ten seconds

Michael Matter, Lena Nenoff, Gabriel Meier, Damien C. Weber, Antony J. Lomax & Francesca Albertini

**To cite this article:** Michael Matter, Lena Nenoff, Gabriel Meier, Damien C. Weber, Antony J. Lomax & Francesca Albertini (2019): Intensity modulated proton therapy plan generation in under ten seconds, Acta Oncologica, DOI: 10.1080/0284186X.2019.1630753

**To link to this article:** https://doi.org/10.1080/0284186X.2019.1630753

© 2019 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

View supplementary material

Published online: 04 Jul 2019.

Submit your article to this journal

Article views: 226

View Crossmark data

Citing articles: 1 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=ionc20

ACTA ONCOLOGICA
https://doi.org/10.1080/0284186X.2019.1630753

Taylor & Francis
Taylor & Francis Group

ORIGINAL ARTICLE

OPEN ACCESS Check for updates

# Intensity modulated proton therapy plan generation in under ten seconds

Michael Matter[a,b], Lena Nenoff[a,b], Gabriel Meier[a], Damien C. Weber[a,c,d], Antony J. Lomax[a,b] and Francesca Albertini[a]

[a]Center for Proton Therapy, Paul Scherrer Institute, Villigen, Switzerland; [b]Department of Physics, ETH Zürich, ETH Hönggerberg, Zürich, Switzerland; [c]Department of Radiation Oncology, University Hospital Zürich, Zürich, Switzerland; [d]Department of Radiation Oncology, University Hospital Bern, Bern, Switzerland

**ABSTRACT**

**Background:** Treatment planning for intensity modulated proton therapy (IMPT) can be significantly improved by reducing the time for plan calculation, facilitating efficient sampling of the large solution space characteristic of IMPT treatments. Additionally, fast plan generation is a key for online adaptive treatments, where the adapted plan needs to be ideally available in a few seconds. However, plan generation is a computationally demanding task and, although dose restoration methods for adaptive therapy have been proposed, computation times remain problematic.

**Material and methods:** IMPT plan generation times were reduced by the development of dedicated graphical processing unit (GPU) kernels for our in-house, clinically validated, dose and optimization algorithms. The kernels were implemented into a coherent system, which performed all steps required for a complete treatment plan generation.

**Results:** Using a single GPU, our fast implementation was able to generate a complete new treatment plan in 5–10 sec for typical IMPT cases, and in under 25 sec for plans to very large volumes such as for cranio-spinal axis irradiations. Although these times did not include the manual input of optimization parameters or a final clinical dose calculation, they included all required computational steps, including reading of CT and beam data. In addition, no compromise was made on plan quality. Target coverage and homogeneity for four patient plans improved (by up to 6%) or remained the same (changes <1%). No worsening of dose-volume parameters of the relevant organs at risk by more than 0.5% was observed.

**Conclusions:** Fast plan generation with a clinically validated dose calculation and optimizer is a promising approach for daily adaptive proton therapy, as well as for automated or highly interactive planning.

**ARTICLE HISTORY**
Received 26 March 2019
Accepted 29 May 2019

## Introduction

Fast generation of treatment plans for proton therapy is important for adapting to major anatomical changes. Such changes are one of the largest source of uncertainty [1] and since manual adaption is labor intensive, much recent research has focused on developing methods for online adaptive proton therapy [2–5]. Ideally, such adaptions should be performed daily with the complete adaption process being performed within 5–10 min. To achieve this, user inputs must be limited, anticipated or their choice automated. In this case, time for adaption will be dominated by plan computation, where calculational performance is of upmost importance. Additionally, fast plan computations will improve the usability of a treatment planning system (TPS), allowing for a more trial and error based planning approach by facilitating efficient searches through the huge solution space of intensity modulated proton therapy (IMPT).

Graphical processing units (GPUs) have been widely deployed for speeding up computationally demanding tasks in medical physics. Different groups have applied them to analytical [6–8] and Monte Carlo (MC) dose calculations [2,9–11], as well as to plan optimization [2,9,12] for proton therapy. The fastest reported values for analytical dose calculations are in the sub-second range. Da Silva et al. achieved analytical dose calculations with a double Gaussian kernel in 0.22 sec [6], whereas reported values for GPU supported MC simulations are in the minute range. Ma et al. report MC dose deposition calculations in 1–15 min depending on the size of the patient case [9] and Botas et al. in 2–7 minutes for head and neck cases of varying sizes [2]. For optimization of pencil beam fluences, Ma et al. report additional optimization times of between a half and seven minutes [9] and Botas et al. between twelve seconds and three minutes [2]. However, the whole computational treatment plan generation does not only consist of dose

**CONTACT** Michael Matter michael.matter@psi.ch Center for Proton Therapy, Paul Scherrer Institute, Villigen, Switzerland

Supplemental data for this article can be accessed here.

This article has been republished with minor changes. These changes do not impact the academic content of the article.

© 2019 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

calculation and optimization. Different preparatory steps like reading the input data or calculation of Bragg peak positions also contribute to the computational cost of plan generation. Therefore, using GPUs coherently for the whole process enables further improvements in the efficiency of IMPT planning.

Alternatively, adaptive planning can instead aim to restore pencil beam positions in the daily patient geometry [13], add pencil beams to cover the target [3] and/or re-optimize pencil beam fluences to restore the nominal dose [4]. Such approaches have the benefit that the adapted is similar to the nominal plan, facilitating clinical validation of the adapted plan. However, dose restoration does not necessarily allow for the best dose distribution on the daily patient geometry, since the new geometry may allow for better target coverage and/or organ at risk (OAR) sparing than in the nominal plan. In addition, calculation times are on the upper border of those required for practical daily adaptive proton therapy.

This study reports on an ultra-fast method for treatment plan generation without compromises as compared to our clinically and well-established treatment planning process. Treatment plan generation times for four example patients are reported, and the quality of the calculated doses evaluated. The benefits of the GPU implementation for online adaptive proton therapy, as well as advantages and drawbacks to other reported methods are discussed. Finally, additional benefits and applications of ultrafast plan generation are described.

## Material and methods

### *Computational plan generation on a GPU*

Our in-house developed TPS uses the ray-casting dose calculation [14], and a quasi-Newtonian fluence optimization [15]. In this work, these algorithms, and other steps of the plan generation process, were implemented onto dedicated Aparapi GPU kernels using Java (Oracle Corporation, Redwood Shores, CA, USA). Aparapi is an opensource framework for executing native Java code on GPUs, based on OpenCL (Khronos Group, Beaverton, OR, USA). The GPU implementation is described in detail in the Supplementary Material. Here we focus on the differences between the clinical and GPU implementation, since these are relevant for interpreting the results.

Although all GPU algorithms were identical to those in our clinical TPS, there were three notable differences in their implementation. First, in the clinical TPS, every field had a set of separate optimization points defined in the beam coordinate system, with the dose being transformed and accumulated to the patient geometry each iteration of the optimization. For the GPU implementation, identical optimization points were used for every field and the coordinate transformations could be dropped, speeding up the optimization and mitigating the need for interpolation. Further, the omission of the interpolation improved the optimization result. Second, the dose distribution of every pencil beam to every optimization point was pre-calculated and stored in a dose deposition matrix we refer to as the $D_{ij}$ matrix. Whilst the GPU implementation used single precision for floating point handling, double precision was used in the clinical implementation. This, however, had a negligible effect on the results. Finally, for efficiency, dose was optimized only in the target and organs at risk (OARs) for which constraints were defined, a region we call the volume relevant for optimization (VRO). For full clinical dose calculations throughout the CT, we used the clinical dose calculation.

### *Patient cases*

Four different patients, previously treated at our institute and spanning a spectrum of indications, were chosen to evaluate the performance of our GPU implementation (Figure 1). Target volume and plan geometries are provided in Table 1. For all, plans were generated with our clinical TPS and then re-generated with the GPU implementation using identical input data and parameters. GPU calculations were performed on a workstation consisting of an Intel Xeon four core 3.5 GHz CPU (v5) and a single Nvidia Quadro P6000 GPU.

### *Comparison between the GPU optimized and the clinical plans*

To investigate the quality of the GPU optimization, plans optimized on the GPU were recalculated with the clinical dose calculation. From these, dose volume histogram (DVH) parameters of the clinical target volume (CTV) and planning target volume (PTV), as well as all relevant OARs were calculated and compared, and differences to the clinically applied plans reported.

### *Comparison to Monte Carlo simulations*

To assess the accuracy of the analytical ray-casting algorithm, all clinical dose plans were compared to full MC simulations. The clinical dose calculation was used for this comparison, because the GPU implementation only provided a dose calculation in the VRO. Differences between the GPU and the clinical implementation however were negligible. MC simulations were performed using TOPAS version 3.0.p1 [16], based on Geant 4 [17] version 10.02.p01, and using the default TOPAS physics list [16,18]. All simulations were performed using the optimized fluence of protons for each field, divided by 1,000, and dose was scored at the resolution of the planning CT. The resulting dose distributions were compared using 3D gamma evaluation (2%/2 mm) considering only voxels with a dose above 10% of the prescription dose and excluding air voxels. A cutoff of 10% was selected to be compatible with values in the literature [19].

## Results

GPU generated plans for all cases are displayed in Figure 1, and times for their complete generation reported in Table 1. For the paraspinal, brain and paranasal patient (3–4 fields,

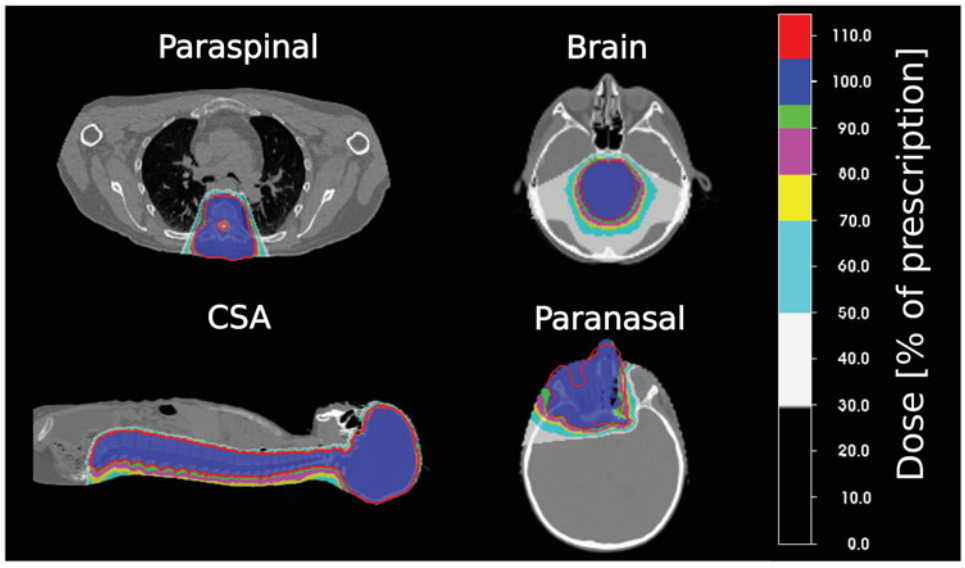


**Figure 1.** Treatment plans generated with the GPU implementation of our in-house developed TPS for four example patient cases: paraspinal, cranio-spinal axis (CSA), brain and paranasal. The planning target volume is outlined in red.

**Table 1.** Treatment plan generation time and plan information for the four patient cases: paraspinal, cranio-spinal axis (CSA), brain and paranasal.

| Case | Paraspinal | CSA | Brain | Paranasal |
|---|---|---|---|---|
| Number of fields | 3 | 2 | 3 | 4 |
| Number of spots | 56,665 | 182,260 | 14,870 | 32,419 |
| Target volume ($cm^3$) | 519 | 2728 | 116 | 195 |
| Number of optimization points | 20,641 | 140,896 | 8411 | 9556 |
| Number of nonzero elements in $D_{ij}$ [1e6] | 94 | 289 | 17 | 81 |
| **Total plan generation time (s)** | 9.9 | 23.5 | 4.6 | 7.4 |
| Preparatory steps calculation time (s) | 4.8 | 6.2 | 3.3 | 3.5 |
| $D_{ij}$ calculation time (s) | 2.3 | 9.5 | 0.6 | 1.6 |
| Pencil beam fluence optimization time (s) | 2.8 | 7.8 | 0.7 | 2.3 |

PTV volumes 116–519 $cm^3$), all plans were generated in <10 s, and for the cranio-spinal axis (CSA) patient (2 fields, volume 2787 $cm^3$) in <25 s. Times for the generation of the $D_{ij}$ matrix and optimization scaled linearly with the number of Bragg peaks and optimization points. In contrast, calculation times required for the preparatory steps (e.g., reading the input data and placing Bragg peaks) depended mainly on the size of the VRO and the number of fields.

Plans optimized with the GPU implementation and the clinically applied plans are compared in Figure 2. Target parameters were improved for 21 out of 24 parameters, where for the remaining three parameters, deterioration was marginal (<1%). Dose–volume parameters of the dosimetrically relevant OARs remained almost identical with no parameter worsening by more than 0.5%. Overall plan quality remained the same or was slightly improved for the plans generated with the GPU implementation. The slight improvement was due to the omission of the interpolations during the optimization.

The ray-casting calculations (as used in the GPU implementation) and TOPAS MC simulations agreed well, with Gamma 2%/2 mm pass rates of 94.3% for the paraspinal, 97.4% for the CSA, 99.8% for the brain and 93.5% for the paranasal patient, respectively.

## Discussion

A method to fully re-generate proton plans in under 10 s has been described for small-to-medium sized tumors, and under 25 s for a cranio-spinal irradiation. This enables online adaptive proton therapy treatments with little deviation from plans generated with the normal clinical workflow and our established and clinically validated TPS.

The times for complete plan generation reported here are considerably shorter than those reported in the literature. The main reason being the use of an analytical dose calculation and a simple, but effective optimization algorithm. Different groups describe MC based proton plan generation on GPUs [2,9,12], with substantially longer times for full plan generation due to the inherently computationally more intense MC approach. However, dose calculation times reported here are similar to those reported for the GPU implementation of an analytical algorithm reported by da Silva et al. [6]. Optimization times are substantially reduced in our work, due to the relatively sparse sampling of optimization points, with other publications sampling at the resolution of the CT grid [2,9,12]. However, we observed no major improvement in results with our GPU implementation when using a finer grid. In addition, by combining all steps into a single coherent, GPU based approach, plan generation times could be further reduced.

In this work, despite using a simple ray-casting dose calculation, agreement to MC simulations is comparable to more widely used pencil beam algorithms. This is interesting, as the ray-casting algorithm uses just a single Gaussian to describe lateral scatter, thus ignoring inelastic and elastic scattering processes. Instead, the field dose is scaled by a boosting factor determined using an empirical model. While this procedure is a major simplification, it has been shown to work well clinically over many cases and match the performance of other analytical calculations reported in the literature [19,20]. Consequently, for applications where calculation time

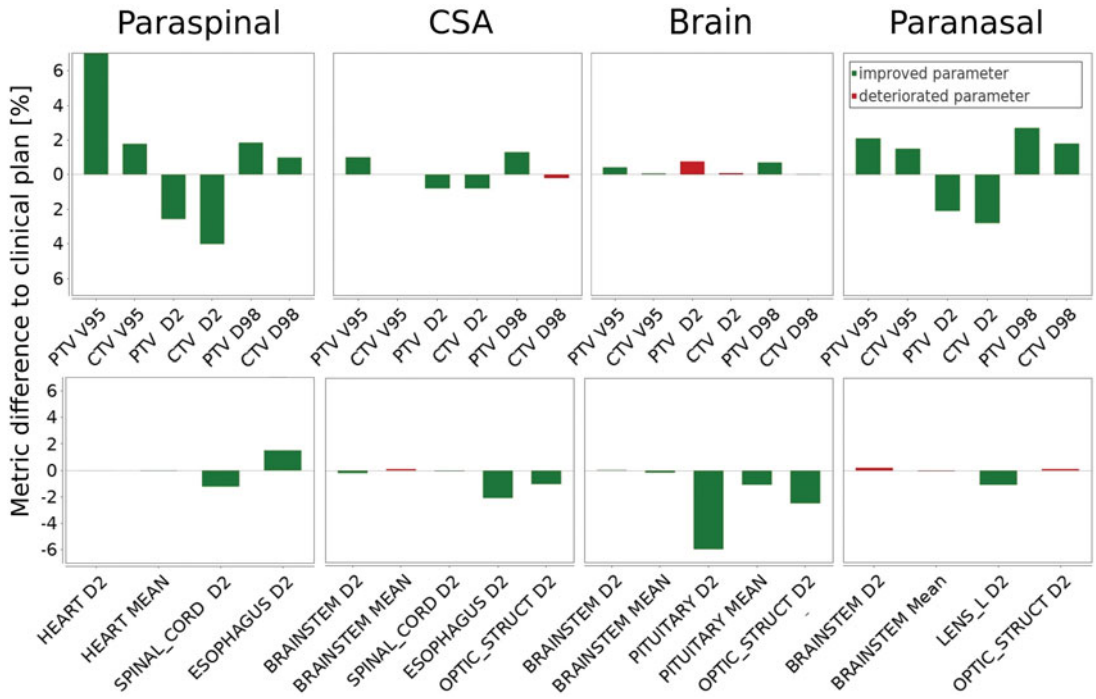


**Figure 2.** Comparison between the DVH parameters of the GPU optimized plans and the clinical plans for the four example cases. In the top row target parameters are displayed and in the bottom row OAR parameters. Improved parameters are depicted in green and deteriorated parameters in red.

is a critical factor (such as for rapid plan adaption on a daily basis), analytical calculations still have a role. For instance, it has been shown elsewhere that anatomical changes affect the dose distribution to a much larger extent than moving from analytical to MC dose calculations [21]. Thus, the use of ultra-fast analytical calculations to enable rapid plan regeneration to mitigate anatomical changes could help solve one of the major uncertainties in proton therapy.

The problem of plan generation for online adaptive proton therapy has recently been addressed by several authors [2–5], where it has been recognized that manual user inputs have to be minimized. As such, Bernatowicz et al. [4], as well as Jagt et al. [5] follow a strategy of dose restoration, where they try to reproduce the dose of the original (nominal) plan on the daily patient geometry, with the aim of simplifying the clinical plan approval process. The drawback however is that the re-optimized plan can never be better than the nominal plan, even if the daily anatomy may allow for improved target coverage or OAR sparing. Alternatively, Botas et al. optimized fluences without restriction to restore the nominal dose distribution, by investigating different restraints on how the pencil beam positions of the nominal plan can be restored on the changed anatomy [2]. Both techniques however require the optimization of various parameters in order to obtain acceptable results.

In contrast, the approach described here is to generate a completely new plan from scratch (using the same field geometry and dose-volume constraints) which is then independent of the nominal plan. For paranasal patients with changes in the nasal cavity, the use of the same field angles and OAR constraint priorities as for the nominal plan, together with complete plan re-generation, has been shown to provide excellent results without the need for dose restoration techniques [22]. Indeed, it has also been shown that improved dose distributions could result due to the sometimes more advantageous anatomical situation encountered in the new patient geometry (e.g., less air/bone interfaces due to increased nasal cavity fillings etc.) With the approach reported here, such advantageous circumstances can be better exploited. Although this may require the development of an automated or simplified plan approval process, we believe this extra effort is vindicated by the potential benefits of the approach.

One limitation in the current GPU implementation is that the dose calculation is limited to the VRO, since only these voxels are relevant for the optimization, whereas for a thorough dose review, the dose in the whole patient should be considered. The GPU implementation however can be extended to calculate the dose over the whole patient geometry at the end of the optimization, which would only slightly increase calculation times (2–4 s). On the other hand, the calculation times quoted here can certainly be reduced more by, for example, optimizing memory transfers to the GPU, the gain of which is difficult to estimate. A further limitation of the described GPU implementation arises from the use of the simple optimization algorithm. With this, full plan automation might be more difficult to achieve in comparison to more complex approaches such as interior point [23] or multi criteria optimization [24].

The speed of this GPU implementation will not only make it useful for adaptive proton therapy applications, but also could have a major impact on the way the TPS will be used for the generation of the initial plan. The computation time of an IMPT plan generation with our clinically used TPS currently takes between a few minutes to an hour for larger volumes. This severely limits the possibility to navigate between

different beam angles and constraint priorities. With the new implementation, a wider spectrum of initial parameters can be tested. For instance, if input parameters or computation settings are changed, only calculations which are downstream of the implied change have to be recalculated. For example, if only the constraint priorities are changed, the $D_{ij}$ matrix and all other starting calculations do not have to be repeated, only the optimization itself. Alternatively, if the beam angle of one field is changed, only the WED map of this field and the corresponding part of the $D_{ij}$ matrix have to be updated, before a new optimization can start. With this approach of only recalculating the data downstream of what has been changed, combined with the fast plan generation times, a new form of highly interactive planning becomes possible, providing a tool for efficiently searching the large solution space of IMPT plans.

In conclusion, ultra-fast calculation times for full plan generation have been achieved with a GPU implementation of a fast analytical dose calculation, such that the time to completely generate treatment plans is no longer a limitation for online adaptive planning. As a clinical implementation of online adaption also requires additional tasks such as contour propagation and plan specific quality assurance, an ultra-fast plan generator provides more time to address these additional tasks. In addition, the approach opens the door to highly interactive IMPT planning, allowing to efficiently search the IMPT solution space.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Funding

The authors would like to acknowledge the Swiss National Science Foundation for their support with the grant project SNF: 165961 *Towards the Daily Adapted Proton Therapy at PSI*, which enabled the making of this publication. The GPU used to conduct the development and the experiments was received from the NVIDIA GPU grant program.

## ORCID

Michael Matter http://orcid.org/0000-0002-2612-3207

## References

[1] Albertini F, Bolsi A, Lomax AJ, et al. Sensitivity of intensity modulated proton therapy plans to changes in patient weight. Radiother Oncol. 2008;86:187–194.

[2] Botas P, Kim J, Winey B, et al. Online adaption approaches for intensity modulated proton therapy for head and neck patients based on cone beam CTs and Monte Carlo simulations. Phys Med Biol. 2018;64:015004.

[3] Jagt T, Breedveld S, van Haveren R, et al. An automated planning strategy for near real-time adaptive proton therapy in prostate cancer. Phys Med Biol. 2018;63:135017.

[4] Bernatowicz K, Geets X, Barragan A, et al. Feasibility of online IMPT adaptation using fast, automatic and robust dose restoration. Phys Med Biol. 2018;63:085018.

[5] Jagt T, Breedveld S, van de Water S, et al. Near real-time automated dose restoration in IMPT to compensate for daily tissue density variations in prostate cancer. Phys Med Biol. 2017;62: 4254–4272.

[6] da Silva J, Ansorge R, Jena R. Fast pencil beam dose calculation for proton therapy using a double-Gaussian beam model. Front Oncol. 2015;5:281.

[7] Fujimoto R, Kurihara T, Nagamine Y. GPU-based fast pencil beam algorithm for proton therapy. Phys Med Biol. 2011;56:1319–1328.

[8] Mein S, Choi K, Kopp B, et al. Fast robust dose calculation on GPU for high-precision 1H, 4He, 12C and 16O ion therapy: the FRoG platform. Sci Rep. 2018;8:14829.

[9] Ma J, Beltran C, Seum Wan Chan Tseung H, et al. A GPU-accelerated and Monte Carlo-based intensity modulated proton therapy optimization system. Med Phys. 2014;41:121707.

[10] Senzacqua M, Schiavi A, Patera V, et al. A fast-Monte Carlo toolkit on GPU for treatment plan dose recalculation in proton therapy. J Phys: Conf Ser. 2017;905:012027.

[11] Qin N, Botas P, Giantsoudi D, et al. Recent developments and comprehensive evaluations of a GPU-based Monte Carlo package for proton therapy. Phys Med Biol. 2016;61:7347–7362.

[12] Qin N, Shen C, Tsai MY, et al. Full Monte Carlo-based biologic treatment plan optimization system for intensity modulated carbon ion therapy on graphics processing unit. Int J Radiat Oncol Biol Phys. 2018;100:235–243.

[13] Zhang M, Westerly D, Mackie T. Introducing an on-line adaptive procedure for prostate image guided intensity modulate proton therapy. Phys Med Biol. 2011;56:4947–4965.

[14] Schaffner B, Pedroni E, Lomax A. Dose calculation models for proton treatment planning using a dynamic beam delivery system: an attempt to include density heterogeneity effects in the analytical dose calculation. Phys Med Biol. 1999;44:27–41.

[15] Lomax A. Intensity modulation methods for proton radiotherapy. Phys Med Biol. 1999;44:185–205.

[16] Perl J, Shin J, Schümann J, et al. TOPAS: an innovative proton Monte Carlo platform for research and clinical applications. Med Phys. 2012;39:6818–6837.

[17] Agostinelli S, Allison J, Amako K, et al. Geant4a simulation toolkit. Nucl Instrum Methods Phys Res A. 2003;506:250–303.

[18] Zacharatou Jarlskog C, Paganetti H. Physics settings for using the Geant4 Toolkit in proton therapy. IEEE Trans Nucl Sci. 2008;55: 1018–1025.

[19] Yepes P, Adair A, Grosshans D, et al. Comparison of Monte Carlo and analytical dose computations for intensity modulated proton therapy. Phys Med Biol. 2018;63:045003.

[20] Winterhalter C, Zepter S, Shim S, et al. Evaluation of the ray-casting analytical algorithm for pencil beam scanning proton therapy. Phys Med Biol. 2019;64:065021.

[21] Nenoff L, Matter M, Geetanjli JA, et al. Anatomical changes vs calculation approximations: which causes larger dose distortions for proton therapy patients? ICCR-MCMA 2019; 2019 June 17–21; Montreal, Canada; 2019.

[22] Nenoff L, Matter M, Hedlund-Lindmar J, et al. Daily adaptive proton therapy: the key to use innovative planning approaches for paranasal cancer treatments. Acta Oncol. 2019; in press.

[23] Breedveld S, Storchi PRM, Voet PWJ, et al. iCycle: integrated, multicriterial beam angle, and profile optimization for generation of coplanar and noncoplanar IMRT plans. Med Phys. 2012;39: 951–963.

[24] Kamal-Sayed H, Ma J, Tseung H, et al. Adaptive method for multicriteria optimization of intensity-modulated proton therapy. Med Phys. 2018;45:5643–5652.