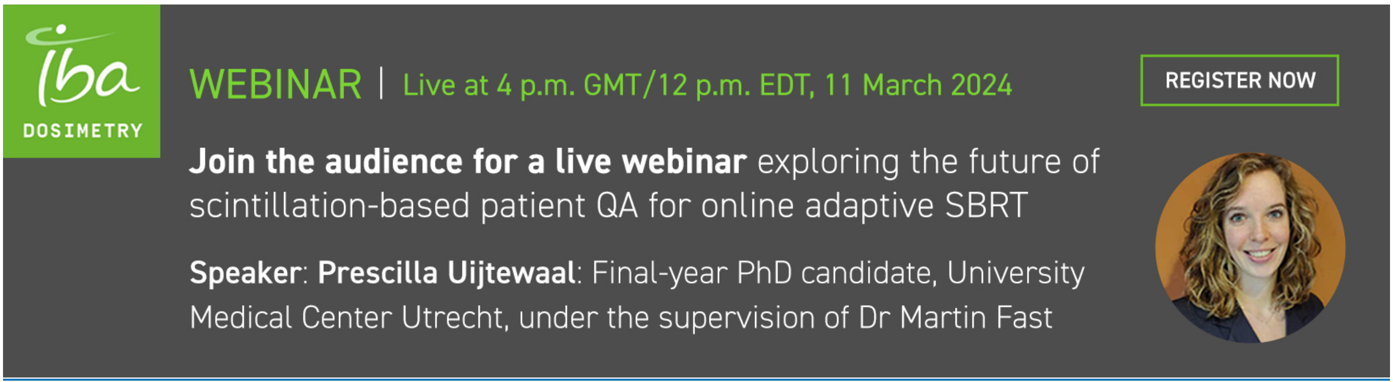
Physics in Medicine & Biology

IPEM
Institute of Physics and
Engineering in Medicine

PAPER • OPEN ACCESS

Evaluation of monte carlo to support commissioning of the treatment planning system of new pencil beam scanning proton therapy facilities

To cite this article: D Botnariuc *et al* 2024 *Phys. Med. Biol.* **69** 045027



IOP Publishing Phys. Med. Biol. **69** (2024) 045027 https://doi.org/10.1088/1361-6560/ad1272

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

RECEIVED
17 July 2023

REVISED
16 November 2023

ACCEPTED FOR PUBLICATION
5 December 2023

PUBLISHED
13 February 2024

PAPER

Evaluation of monte carlo to support commissioning of the treatment planning system of new pencil beam scanning proton therapy facilities

D Botnariuc[1,2,*], **S Court**[3], **A Lourenço**[1,2], **A Gosling**[3], **G Royle**[1], **M Hussein**[2], **V Rompokos**[3] **and C Veiga**[1]

[1] Department of Medical Physics and Biomedical Engineering, University College London, Gower Street, London, WC1E 6BT, United Kingdom

[2] Metrology for Medical Physics Centre, National Physical Laboratory, Hampton Road, Teddington, TW11 0LW, United Kingdom

[3] Radiotherapy Physics Services, University College London Hospitals NHS Foundation Trust, 250 Euston Road, London, NW1 2PG, United Kingdom

[*] Author to whom any correspondence should be addressed.

E-mail: daniela.botnariuc.19@ucl.ac.uk

Keywords: proton therapy, monte carlo, beam modelling, commissioning

Supplementary material for this article is available online

ABSTRACT

Objective. To demonstrate the potential of Monte Carlo (MC) to support the resource-intensive measurements that comprise the commissioning of the treatment planning system (TPS) of new proton therapy facilities. *Approach.* Beam models of a pencil beam scanning system (Varian ProBeam) were developed in GATE (v8.2), Eclipse proton convolution superposition algorithm (v16.1, Varian Medical Systems) and RayStation MC (v12.0.100.0, RaySearch Laboratories), using the beam commissioning data. All models were first benchmarked against the same commissioning data and validated on seven spread-out Bragg peak (SOBP) plans. Then, we explored the use of MC to optimise dose calculation parameters, fully understand the performance and limitations of TPS in homogeneous fields and support the development of patient-specific quality assurance (PSQA) processes. We compared the dose calculations of the TPSs against measurements ($DD_{\text{TPSvs.Meas.}}$) or GATE ($DD_{\text{TPSvs.GATE}}$) for an extensive set of plans of varying complexity. This included homogeneous plans with varying field-size, range, width, and range-shifters (RSs) ($n = 46$) and PSQA plans for different anatomical sites ($n = 11$). *Main results.* The three beam models showed good agreement against the commissioning data, and dose differences of 3.5% and 5% were found for SOBP plans without and with RSs, respectively. $DD_{\text{TPSvs.Meas.}}$ and $DD_{\text{TPSvs.GATE}}$ were correlated in most scenarios. In homogeneous fields the Pearson's correlation coefficient was 0.92 and 0.68 for Eclipse and RayStation, respectively. The standard deviation of the differences between GATE and measurements (±0.5% for homogeneous and ±0.8% for PSQA plans) was applied as tolerance when comparing TPSs with GATE. 72% and 60% of the plans were within the GATE predicted dose difference for both TPSs, for homogeneous and PSQA cases, respectively. *Significance.* Developing and validating a MC beam model early on into the commissioning of new proton therapy facilities can support the validation of the TPS and facilitate comprehensive investigation of its capabilities and limitations.

1. Introduction

Proton beam therapy (PBT) is becoming increasingly available for cancer treatment worldwide. The growing interest in making PBT more available reflects the favourable dose distributions that can be achieved, allowing for integral dose reduction in younger patients and the treatment of complex diseases (Foote *et al* 2012, Mohan 2022). With a remarkable number of facilities currently treating patients and many more in development stages (PTCOG 2023), efficient clinical commissioning standards and procedures become more critical. Commissioning of the multiple individual components of the proton therapy system is essential prior to clinical use. These include the treatment delivery system (accelerator, beamline and nozzle), the patient imaging components and the treatment planning system (TPS), amongst others (Farr *et al* 2021).

The TPS is one of the key elements of the PBT system as it is used to plan, optimise, and assess patients' treatments. The clinical commissioning of the TPS includes beam data acquisition, modelling of the beam and validation of the calculated doses, for which a set of comprehensive but time-consuming experimental measurements are required. For the TPS dose validation in pencil beam scanning (PBS) systems, the American Association of Physicists in Medicine (AAPM) Task Group 185 recommends measurements of dose outputs in both uniform and non-uniform fields of varying complexity (Farr *et al* 2021). The uniform fields in homogeneous media should comprise monoenergetic layers and spread-out Bragg peaks (SOBPs) of different combinations of field size, range, and width (modulation). The non-uniform fields consist of a variety of patient-specific quality assurance (PSQA) plans, representative of clinical indications to be treated at the facility. Both types of fields should be verified for plans with and without beam modifying devices like range shifters (RSs) (Farr *et al* 2021). It is fundamental that the dose calculation algorithms used clinically are accurate and this must be assessed during commissioning, prior to starting patient treatments.

Analytical pencil beam (PB) algorithms embedded within TPSs are the most popular tools for dose calculations in the clinic. The main limitation of analytical models is the poor modelling of the lateral scatter which leads to inaccurate dose calculation in complex and heterogeneous media (de Martino *et al* 2021, Saini *et al* 2018). Monte Carlo (MC) codes are considered the most accurate dose calculation engines in radiotherapy, as these simulate particle transport in matter based on fundamental particle interactions (Paganetti *et al* 2008, Paganetti 2012, Grassberger *et al* 2015, Yepes *et al* 2018, Tommasino *et al* 2018). Historically, MC algorithms were available in very specialised software packages tailored for research use, limiting their application in clinical settings (Waters *et al* 2007, Agostinelli *et al* 2003, Böhlen *et al* 2014). More recently, there has been an effort to develop toolkits to make general-purpose MC codes more user-friendly for researchers and clinical staff working in medical physics applications. Examples of such toolkits are TOPAS (Perl *et al* 2012) and GATE (Jan *et al* 2011, Sarrut *et al* 2014) for Geant4 (Agostinelli *et al* 2003) and Flair (Vlachoudis 2009) for FLUKA (Böhlen *et al* 2014). Similarly, commercial TPSs have started to incorporate MC algorithms for dose calculations—for example, AcurosPT in Eclipse (Varian Medical Systems) and the MC algorithm in RayStation (RaySearch Laboratories). To reduce computational times, strategies like simplifying or neglecting some particle transport processes (Schreuder *et al* 2019, Varian Medical Systems 2020) and/or GPU (Graphical Processing Units) implementations (Varian Medical Systems 2023, RaySearch Laboratories 2023) are often employed. Regardless of the superiority of MC, PB algorithms are still largely used clinically due to their convenience and availability.

The simulation of clinical treatment plans in general purpose MC toolkits requires accurate modelling of the incident beam. MC models of clinical beams may be used to support decision making at clinical facilities. The most popular use of general purpose MC codes in proton therapy is for independent dose calculations, with multiple studies demonstrating in-house workflows for this application (Paganetti *et al* 2008, Tourovsky *et al* 2005, Grevillot *et al* 2012, Magro *et al* 2015, Fracchiolla *et al* 2015, Aitkenhead *et al* 2020, Verburg *et al* 2016, Guterres Marmitt *et al* 2020). However, there are other applications that have been less explored. The AAPM Task Group 185 suggests MC as an alternative to direct measurements to support the different stages of the commissioning process of new proton facilities (Farr *et al* 2021). In this context, some studies have shown that MC generated beam data could reduce the number of measurements required to configure a beam model in the TPS (Newhauser *et al* 2007, Clasie *et al* 2012). Alternatively, an adequately validated MC model of a PBS system could be developed early into the commissioning stage of a new facility to support the dose validation of the radiotherapy TPS. This could help reduce the number of TPS validation measurements performed during this process (for example, to specific field configurations where MC would indicate larger dosimetric differences), enhance the number and variety of cases tested, as well as allow a more comprehensive understanding of the dose calculation engine and its limitations before starting the treatment of patients. The number of measurements required during commissioning of new facilities are resource- and time-consuming. It is recognised that efficiency improvements during the early stages of a new facility can lead to transitioning to the routine clinical phase on schedule and reduce the risk of delays going clinical.

In this work, our aim was to demonstrate the potential of MC during the TPS dose validation process for a new proton therapy facility. First, we tuned and benchmarked a model of the proton system installed in our institution using beam commissioning data and a limited number of homogeneous fields. Then, with a properly benchmarked MC implementation, we demonstrated the potential of MC to complement the extensive measurements that comprise the optimisation and validation of the TPS of new proton therapy facilities. The analysis was performed for two TPSs to validate our approach for multiple dose calculation engines. To the best of our knowledge, this is the first time an independent MC tool was investigated in detail to support the TPS dose verification and validation during the commissioning of a new PBS system.

2. Methods and materials

The proton beam system modelled in this work was the Varian ProBeam (Varian Medical Systems) installed at University College London Hospitals proton centre, in the UK. This clinical system was retrospectively modelled in GATE (Jan *et al* 2011, Sarrut *et al* 2014), an open-source MC package, and in two commercial TPS systems—Eclipse (Varian Medical Systems) and RayStation (RaySearch Laboratories).

2.1. Beam commissioning data

The beam commissioning data used consisted of measurements of integral depth-doses (IDDs) and absolute dose calibration in water and spot profiles in air, for nominal energies ranging from 70 to 245 MeV, in steps of 5 MeV. The IDDs of monoenergetic pencil beams were measured in a water phantom, using a 4.08 cm sensitive radius plane-parallel PTW Bragg peak chamber 34070 (PTW-Freiburg). The absolute dose was determined by measuring the dose at the reference depth of 2 cm water-equivalent depth, for 10 × 10 cm^2 layers, with 2.5 mm spot spacing, using a plane-parallel PTW Roos chamber 34001 (PTW-Freiburg). For each IDD, the whole curve was normalised to the absolute dose determined at the reference depth. Finally, the IDDs were scaled according to a relative biological effect (RBE) factor of 1.1 [units: Gy mm^2 MU^{-1} (RBE)]. The spot profiles were measured, both with and without RSs in the beam's path, at the isocentre and at six distances both up and downstream from the isocentre plane (0 cm, ±10 cm, ±20 cm and ±40 cm), using the 42 × 32 cm scintillation detector XRV-4000 Hawk Beam Profiler (Logos Systems). The uncertainty of the measured profiles is within ± 0.1 mm at the full width at half maximum position, and within ±0.3 mm at centroid positions. The spot sizes in the x and y directions according to the beam's eye view coordinates were defined as one standard deviation (σ) of a Gaussian function fitted to the measured spot profiles. For spot profile measurements with each of the three RS options (40 × 30 cm^2 Lexan Polycarbonate blocks of 2, 3 and 5 cm thickness), the RSs were inserted into a fixed position within the ProBeam's fully retracted snout. The water equivalent thickness (WET) of each RS was measured using the Giraffe detector (Ion Beam Applications SA), which has an uncertainty in range determination of 0.5 mm.

2.2. Beam modelling and dose calculation in Eclipse and RayStation

Beam models were built for the proton convolution superposition (PCS) algorithm in Eclipse v16.1 and the MC algorithm in RayStation v12.0.100.0 following the specifications provided by the manufacturers (Varian Medical Systems 2020, RaySearch Laboratories 2019). In Eclipse, the user must provide measurements of IDDs (calibrated to units of Gy mm^2 MU^{-1} (RBE)) and spot profiles both with and without RSs. The PCS algorithm requires measurement data of spot profiles with RSs to model the lateral scattering of the beam in the presence of RSs. The lateral cut-off calculation parameter, σ_{Ecl}, was set to the maximum value of 4 (unless stated otherwise). In RayStation, the user must provide measurements of IDDs, absolute dose calibration and spot profiles without RSs only, as it is not an option to import spot profiles measured with RSs. The material of the RSs was selected according to the details provided by the manufacturer.

2.3. Beam modelling in GATE

The MC model of the clinical beam was implemented using GATE v8.2. Our modelling approach consisted of a parameterisation of the pencil beam source as a function of the nominal energy. The pencil beam type source in GATE is characterised by its energy (E_{mean} – mean energy and σ_E – energy spread) and optical parameters (σ_x and σ_y – spot size, σ_θ andσ_φ – spot divergence and ϵ_x and ϵ_y – emittance in the x and y directions, respectively). The source was positioned beyond the nozzle exit, prior to the beam modifying devices, at 60 cm upstream from the isocentre, so that the RSs could be physically modelled in the simulations. The parametrisation was achieved by tuning iteratively the beam properties at the source to match the experimental beam commissioning data in the absence of any RSs. The RSs were then modelled at the nozzle exit as 40×30 cm^2 blocks of Lexan Polycarbonate (fraction by weight: 5.5% H, 75.5% C, 19% O) with varying thicknesses (2, 3 and 5 cm). The density of Lexan Polycarbonate is 1.21 g cm^{-3} according to the manufacturer specifications (Varian Medical Systems 2020); this density was empirically adjusted to 1.195 g cm^{-3} in the MC simulations to match the simulated and experimental WET.

Figure 1 shows a summary of the beam parametrisation methodology, which consisted of three sequential optimisation steps: (1) optical parameters, (2) energy parameters and (3) absolute dose calibration. The fraction of energy scored during the optimisation of the IDDs is dependent on the optical properties (Grevillot *et al* 2011). Therefore, the optical parameters were determined first and were used in the tuning of the energy parameters (Grevillot *et al* 2011, Yeom *et al* 2020). The absolute dose calibration was obtained by normalising the area under the curve of the IDDs in GATE to that of the measured IDDs. The QGSP_BIC_EMZ physics list was used in all simulations and the production cuts on secondary particles (electrons, photons, positrons) were

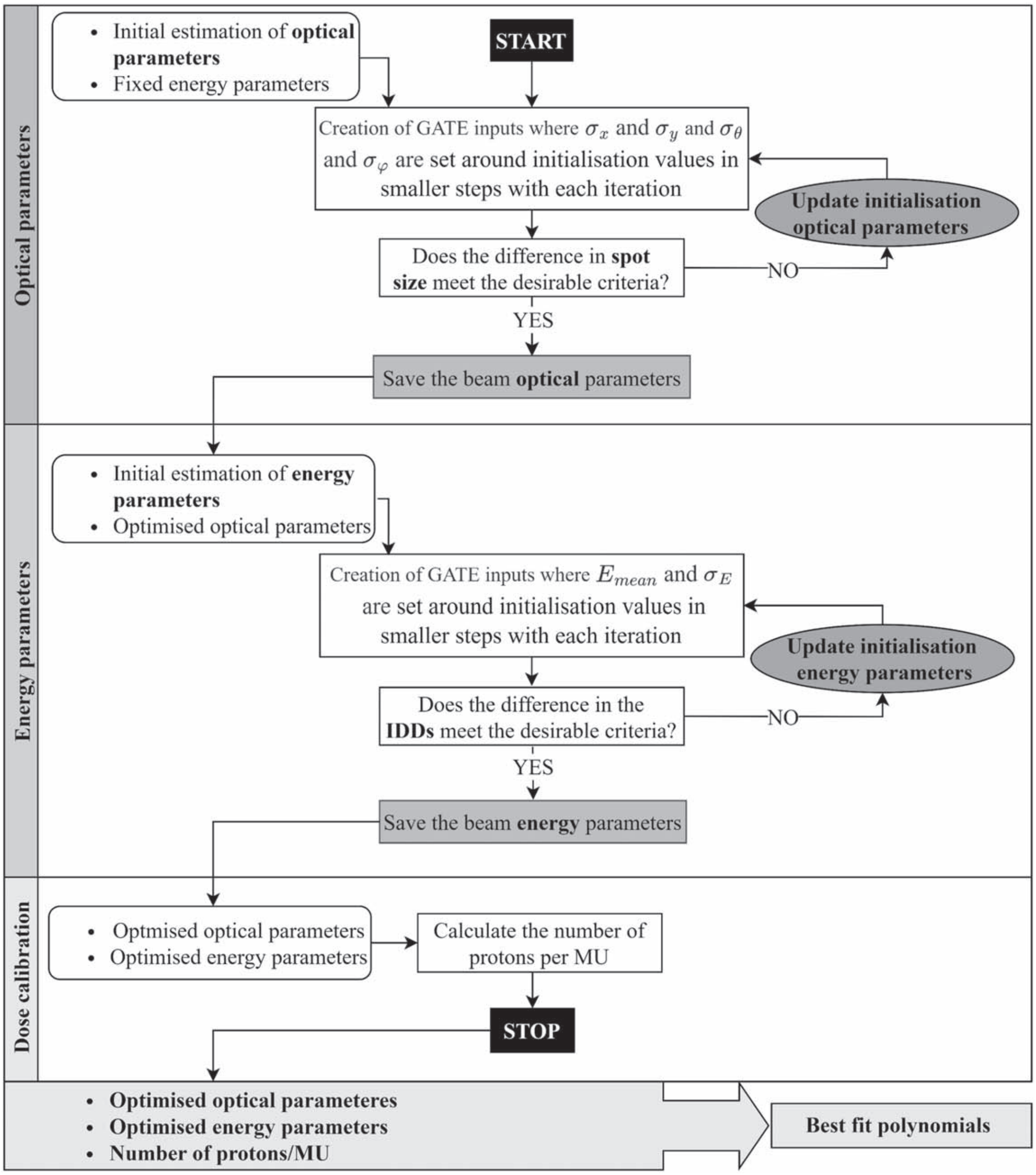

Optical parameters
Initial estimation of optical parameters
Fixed energy parameters
START
Creation of GATE inputs where σ_x and σ_y and σ_θ and σ_φ are set around initialisation values in smaller steps with each iteration
Update initialisation optical parameters
Does the difference in spot size meet the desirable criteria?
NO
YES
Save the beam optical parameters
Energy parameters
Initial estimation of energy parameters
Optimised optical parameters
Creation of GATE inputs where E_{mean} and σ_E are set around initialisation values in smaller steps with each iteration
Update initialisation energy parameters
Does the difference in the IDDs meet the desirable criteria?
NO
YES
Save the beam energy parameters
Dose calibration
Optmised optical parameters
Optimised energy parameters
Calculate the number of protons per MU
STOP
Optimised optical parameteres
Optimised energy parameters
Number of protons/MU
Best fit polynomials

Figure 1. Workflow of the automated process developed to model a pencil beam scanning system in GATE.

set to 0.01 mm according to the recommendation provided by Winterhalter *et al* (2020b). All simulations had a statistical uncertainty below 0.25%. The optimisation process was implemented in MATLAB 2019b (Mathworks). Complete details of the optimisation strategy can be found in supplementary material 1.

2.4. Dose calculation parameters for tested plans

All the plans used in this work to demonstrate the potential of MC to complement experimental measurements were optimised in Eclipse using the non-linear universal proton optimizer (NUPO) algorithm (Varian Medical Systems 2020, Nocedal and Wright 2006). Two types of plans were evaluated: homogeneous (such as layers and box-fields) and non-homogeneous (clinical PSQA fields). The doses were calculated using a $40\times40\times40$ cm^3 water box (for homogeneous fields) or a $30 \times 30 \times 30$ cm^3 solid water (PTW RW3) box (for non-homogeneous fields). The water material was defined by assigning to the box structure a relative stopping power (RSP) of 1 for dose calculations. The RSP for the solid water phantom was calculated for multiple solid water slabs as the ratio between the WET, measured with the Giraffe detector (Ion Beam Applications SA), and the physical thickness, and the average RSP of 1.041 ± 0.002 was considered. A resolution of $1 \times 1 \times 1$ mm^3 was the default value used for 3D dose calculations in all algorithms. The plans were then delivered experimentally with the ProBeam at our institution, as well as recalculated independently in the other dose calculation engines—GATE and RayStation.

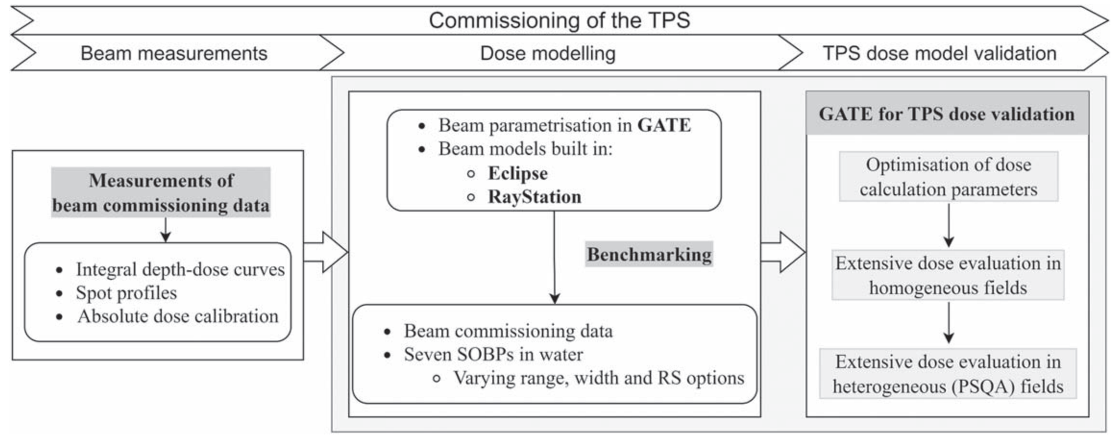

Commissioning of the TPS
Beam measurements
Dose modelling
TPS dose model validation
Measurements of beam commissioning data
Integral depth-dose curves
Spot profiles
Absolute dose calibration
Beam parametrisation in GATE
Beam models built in:
Eclipse
RayStation
Benchmarking
Beam commissioning data
Seven SOBPs in water
Varying range, width and RS options
GATE for TPS dose validation
Optimisation of dose calculation parameters
Extensive dose evaluation in homogeneous fields
Extensive dose evaluation in heterogeneous (PSQA) fields

Figure 2. Workflow of the steps and experiments performed to demonstrate the potential of an independent MC beam model to inform and complement the commissioning and dose validation of the TPS of a new proton therapy facility.

In GATE, the parameterised beam model described in section 2.3 was utilised. For plans recalculated in water, a water box geometry was defined, whereas for plans recalculated in solid water, an image with the corresponding HUs was imported and positioned according to the plan. The water material was defined with a density of 1 g cm^{-3} and mean excitation energy of 78 eV (ICRU report 90 2016). The solid water material was defined as per specifications of the manufacturer (fraction by weight: 7.6% H, 90.4% C, 1.2% Ti, 0.8% O) and its density of 1.045 g cm^{-3} was tuned to 1.057 g cm^{-3} in GATE to match the experimental RSP. In all GATE simulations the number of primary particles simulated was chosen to achieve a statistical uncertainty below 0.5%. In RayStation, the doses were recalculated using the MC algorithm. Water was defined with a density of 1 g cm^{-3} and solid water (same elemental composition as in GATE) was defined with a density of 1.062 g cm^{-3} to match the experimental RSP. All plans were calculated in RayStation with a statistical uncertainty of 0.5%.

2.5. Demonstration of the potential of MC to support TPS commissioning

The experiments performed to demonstrate the potential of MC to support and complement the experimental measurements necessary during the TPS dose validation are summarised in figure 2 and described in detail in the following sections.

2.5.1. Benchmarking of the beam models

The first experiment consisted of benchmarking the beam models built in GATE and the two TPSs against experimental measurements with the aim of verifying the accuracy of their implementation and performance in a limited but representative number of scenarios. This step should be performed in the early stage of the commissioning to gain confidence in both the MC and the TPS's models. Therefore, the benchmarking data included the beam commissioning data and a limited number of IDDs and lateral profiles in SOBPs in water (with and without RSs).

The three beam models were first compared against the beam commissioning data. The AAPM task group 224 recommends range tolerances for the IDDs to be within ± 1 mm and maximum differences of ±10% for the spot sizes (Arjomandy *et al* 2019). In GATE, IDDs and spot profiles (with and without RSs) were simulated using the parametrised beam model. In Eclipse, the calculated IDDs and spot profiles with and without RSs were exported from the system. In RayStation, the model fitted IDDs and spot profiles without RSs were exported from the system, similarly to Eclipse. However, since RayStation does not model range-shifting devices using measurement data, the modelled spot profiles with RSs (2, 3 and 5 cm) could not be directly exported. Instead, treatment plans of single spots in air with RSs of varying thickness for all energies were created. The 3D dose was calculated and the array of voxels in the *x* and *y* directions corresponding to the depth of measurements were extracted from the 3D dose files, using an in-house developed MATLAB code.

The performance of the models was then comprehensively assessed in a set of seven box fields delivered to water: three 10×10 cm^2 spread-out Bragg peaks (SOBP) plans of 15, 20 and 30 cm range (*R*) and 10 cm width (*W*) (without a RS), and four 5×5 cm^2 SOBP plans of 12 cm *R* and 5 cm *W* (one without a RS and three with RSs of varying thickness). For each plan, the following data was measured experimentally: IDDs in the centre of the volumes using a PTW Roos 34001 ionisation chamber and lateral profiles using the PTW microDiamond 60019 detector. IDDs and lateral profiles were extracted from the 3D dose distributions by integrating the dose along the depth for the area of the PTW Roos ionisation chamber and by extracting the lateral array of voxels at the central plane, respectively. Each lateral profile, both measured and calculated, was normalised to its value at the central coordinate.

2.5.2. MC to support TPS commissioning

The second experiment consisted of demonstrating the potential of MC to complement the time-consuming and resource-intensive measurements of TPS dose validation in new facilities. Here we aim to investigate how centres may incorporate an adequately validated MC beam model to support the verification process of their TPS. We investigated how MC may inform the optimisation of TPS calculation parameters (in Eclipse) and the validation of the two TPS models in both homogeneous and non-homogeneous field plans. To demonstrate these application scenarios, the output of the two commercial TPSs was compared against measurements or GATE for an extensive set of plans with diverse complexity just like those used during commissioning to evaluate the TPS dose calculation. A summary of the scenarios and a description of their measurement and calculation in the three algorithms is presented in table 1; complete details are provided in the paragraphs below.

2.5.2.1. Scenario 1: optimisation of calculation parameters—lateral cut-off and dose grid resolution

Commercial dose calculation algorithms may provide the freedom for the user to select their preferred dose calculation parameters. The choice of the dose calculation parameters has a non-negligible influence on the dose calculation accuracy (Zhao 2013). Therefore, as part of commissioning, these parameters must be evaluated to identify the adequate balance between calculation accuracy and computation time. In Eclipse, we investigated two calculation parameters: the lateral cut-off, σ_{Ecl}, and the grid resolution. The σ_{Ecl} calculation parameter is defined as '*the cut-off value for the extent of the lateral dose calculation in units of the beamlet sigma or spot sigma*' (Varian Medical Systems 2020) and may influence the absolute dose calculation. The resolution of the dose grid may influence the local dose distribution significantly, especially for small and non-homogeneous fields (Zhao 2013). In order to demonstrate that these calculation parameters could be tuned relying on MC only, a set of ten 10 × 10 cm^2 SOBP plans in water with varying range and width were simulated in GATE and calculated in Eclipse using a σ_{Ecl} value of 2, 3 and 4. To evaluate the influence of the grid resolution on both homogeneous and non-homogeneous fields, the same ten SOBP plans and seven PSQA cases were recalculated in Eclipse with varying grid resolution (1 × 1 × 1, 2 × 2 × 2 and 3 × 3 × 3 mm^3), using the previously optimised σ_{Ecl} parameter. The point dose outputs at the centre of the SOBP for each parameter from Eclipse were compared to both measurements and GATE, in terms of the mean difference value and standard deviation.

2.5.2.2. Scenario 2: dose evaluation in homogeneous fields

One step in the performance verification of a clinical TPS consists of evaluating the lateral effect of the beam halo by analysing the dose outputs at the centre of squared mono-energetic fields. The Gaussian fit approximation of the transverse spot profiles in analytical dose calculation algorithms disregards the broad tails of the spot profiles, also known as beam halo (Harms *et al* 2020). The beam halo is the result of particle scattering within the beam line and nuclear interactions, which have a larger weight for higher energy beams (Sawakuchi *et al* 2010, Gottschalk *et al* 2015). It is challenging to assess the effect of the beam halo in single spots. However, the cumulative contribution of the low-dose envelope is significant for larger fields, where there is superposition of single pencil beams (Sawakuchi *et al* 2010, Grevillot *et al* 2011, Harms *et al* 2020). The dose outputs for mono-energetic layer fields should increase with increasing field size, reaching a dose plateau when charged particle equilibrium is achieved - as the number of single beams is larger, a larger contribution of low dose is expected in total (Grevillot *et al* 2011). To evaluate the lateral effect of the beam halo, the dose at 2 cm WET for a 100 MeV mono-energetic layer was measured and obtained with the three calculation algorithms for seven different field sizes (3 × 3, 4 × 4, 5 × 5, 7 × 7, 10 × 10, 12 × 12, 15 × 15 cm^2). The 2 cm depth was chosen as it presents a low dose gradient, which decreases detector positioning uncertainties. All dose values were normalised to the output of the 10×10 cm^2 field. The calculated point dose outputs using GATE and both TPS were compared against measurements to evaluate the accuracy of each algorithm. Eclipse and RayStation outputs were also compared against GATE. The contribution of the beam halo for spots across multiple energy layers can be further evaluated in SOBP plans, which will be dependent on SOBP range and width.

An extensive set of uniform SOBP fields of varying field size and varying range/width, covering a representative range of clinical energies, should be evaluated. The AAPM Task Group 185 suggested a list of fields to investigate in table V of their publication (Farr *et al* 2021). Considering the significant number of fields recommended and the associated resources and time required for measurements, a representative subset of plans could be first measured. One could make use of this experimental data to infer the dose outputs for the remaining plans through MC simulations, to gain a better understanding of the TPS dose calculation algorithm and its limitations. If large deviations in dose for these plans are observed when comparing to MC, then those cases should be verified against measurements. A total of 39, 10×10 cm^2 SOBP fields, with varying range (10–35 cm), width (2–20 cm) and range shifter options were tested. Within these, a subset of 10 SOBP fields were

Table 1. Summary of scenarios investigated to demonstrate the potential of Monte Carlo to support the commissioning of clinical treatment planning systems.

Scenario	Plan type	Measurement	GATE		Eclipse		RayStation	
Calculation parameters	Lateral cut-off value SOBP plans with varying range and width in water ($n = 10$)	Point doses in the centre of SOBP (R-W/2) using a PTW Roos 34001 ionisation chamber and point doses per selected PSQA plane using a PTW Semiflex ionisation chamber	$40 \times 40 \times 40$ cm^3 water box; dose scored in a cylinder (radius = 0.78 cm and thickness = 0.1 mm) positioned at R-W/2 of the SOBP		$\sigma_{\mathrm{Ecl}} = 2$, 3 or 4 and resolution of $1 \times 1 \times 1$ mm^3; Dose extracted at the R-W/2 coordinate		N/A	
	Dose calculation resolution SOBP plans with varying range and width in water ($n = 10$) and PSQA plans ($n = 6$)				$\sigma_{\mathrm{Ecl}} = 4$ and resolution of $1 \times 1 \times 1$, $2 \times 2 \times 2$ and $3 \times 3 \times 3$ mm^3; Dose extracted at the R-W/2 coordinate		N/A	
Homogeneous fields	100 MeV mono-energetic layer plans of different field sizes (3×3, 4×4, 5×5, 7×7, 10×10, 12×12, 15×15) delivered to solid water ($n = 7$)	Point doses at 2 cm WET using a PTW Roos 34001 ionisation chamber	$40 \times 40 \times 40$ cm^3 solid water box; dose scored in a cylinder (radius = 0.78 cm and thickness = 0.1 mm) positioned at 2 cm WET		$\sigma_{\mathrm{Ecl}} = 4$ and resolution of $1 \times 1 \times 1$ mm^3	Dose extracted at 2 cm WET in a structure representing the PTW Roos ionisation chamber	Resolution of $1 \times 1 \times 1$ mm^3	Dose extracted at 2 cm WET in a structure representing the PTW Roos ionisation chamber
	SOBP plans with varying range (10 cm to 35 cm), width (2 cm to 20 cm) and range-shifter options delivered to water ($n = 39$)	Point doses in the centre of SOBP (R-W/2) using a PTW Roos 34001 ionisation chamber	$40 \times 40 \times 40$ cm^3 water box; dose scored in a cylinder (radius = 0.78 cm and thickness = 0.1 mm) positioned at R-W/2 of the SOBP			Dose extracted at the R-W/2 coordinate of the SOBP		Dose extracted at the R-W/2 coordinate of the SOBP
Heterogeneous fields	PSQA plans of different anatomical sites delivered to solid water ($n = 11$)	2 to 3 planes per field using a PTW OCTAVIUS array detector 1500XDR	$40 \times 40 \times 40$ cm^3 phantom extracted from the TPS and imported as CT image; dose scored within a $1 \times 1 \times 1$ mm^3 grid	Point doses extracted from the 3D dose map using an in-house script	$\sigma_{\mathrm{Ecl}} = 4$ and resolution of $1 \times 1 \times 1$ mm^3	Point doses extracted from the 3D dose map using an in-house script	Resolution of $1 \times 1 \times 1$ mm^3	Point doses extracted from the 3D dose map using an in-house script
		One point dose per selected plane using a PTW Semiflex 3D 31021 ionisation chamber		Matching plane closest to measurement depth selected from the 3D dose map		Matching plane closest to measurement depth selected from the 3D dose map		Matching plane closest to measurement depth selected from the 3D dose map

R = range, W = Width, SOBP = Spread Out Bragg Peak selected (the same ones as tested for the optimisation of the calculation parameters), ensuring variability in ranges and widths. We then investigated if the standard deviation of the differences between GATE and measurements could be used as threshold to consider when comparing TPS dose outputs with GATE, to identify plans that fail the established passing criteria of 2% (equation (1)). Equation (1) and the threshold value found considering the 10 fields was further applied to the full set of 39 SOBPs.

$$\text{Differences}_{(\text{TPS vs.GATE})} \pm \text{Threshold} < \text{Passing}\ \ \text{Criteria} \tag{1}$$

2.5.2.3. Scenario 3: dose evaluation in non-homogeneous field

The final verification of a clinical TPS before approving its clinical use consists of evaluating dose calculations for a range of preclinical PSQA plans against measurements. This step is essential to understand the behaviour of the dose calculation algorithm for different types of non-homogeneous plans and anatomical sites. Furthermore, standard PSQA procedure workflows should be developed and adopted by the centre, which may also include the use of MC as an independent dose calculation tool. Measuring a range of clinical plans allows one to understand the performance of both the TPS and MC in complex fields.

A total of 11 PSQA plans of different anatomical sites, delivered to solid water, were tested. The anatomical sites included brain, spine, pelvis, head and neck and breast cases. 2D plane doses were measured at up to three pre-defined depths in the solid water phantom per field, using the ionisation chamber matrix PTW OCTAVIUS detector 1500XDR. This detector contains 1405 ionisation chambers ($4.4 \times 4.4 \times 3$ mm^3) arranged on a 27×27 cm^2 matrix, with a centre-to-centre distance between each ionisation chamber of 7.1 mm. Then, for each plane, a dose point was also measured using a PTW Semiflex 3D 31021 cylindrical ionisation chamber. This small ionisation chamber (sensitive volume of 0.07 cm^3) provides high spatial resolution and is adequate for point dose measurements. A total of 72 planes and correspondent points were measured and analysed. All fields were simulated in GATE and calculated in both TPSs, and a 3D dose grid was obtained for each field. The point doses were derived from each 3D dose distribution by taking the coordinates of the measuring points and extracting the dose in a region of interest corresponding to the volume of the PTW Semiflex ionisation chamber. Similar to the SOBP plans, the standard deviation of differences between GATE and measurements for all fields considered was used as threshold when comparing TPS dose outputs with GATE. Point dose differences below 3% were considered as the clinical passing criteria. A 3%/3 mm local gamma analysis with a lower cut dose threshold of 5% was adopted, following our institution's protocol to evaluate PSQA plans. This analysis was performed for the 2D dose planes to test the agreement between the three dose calculation algorithms and the measured arrays, following Hussein *et al* 2017. Gamma pass rates above 95% were defined as within clinically passing criteria (Farr *et al* 2021).

3. Results

3.1. Benchmarking of the beam models

Table 2 summarises the validation data of the beam models against commissioning beam data of IDDs and air profiles, plus a representative set of seven box fields in water. Examples of measured and calculated IDDs and spot profiles for 70 MeV and 245 MeV are shown in supplementary material 2. Overall, the three beam models matched the commissioning beam data within the tolerances described in section 2.5.1 for most cases. However, the Monte Carlo-based beam models (GATE and RayStation) presented a superior performance on average when considering the IDDs and lateral profiles for the tested box field scenarios.

The IDDs modelled in GATE, Eclipse and RayStation were compared against the experimental IDDs in terms of $R_{80\%}$, $W_{80\%}$, area under the curve (AUC), dose at 2 cm depth and dose at the peak. The absolute differences in $R_{80\%}$ (figure 3(a)) were within 0.1 mm for the TPSs and 0.4 mm for GATE. GATE $R_{80\%}$ differences were larger and had a larger standard deviation across the energy range, likely because the energy parameters were optimised with the trade-off of balancing different quantities ($R_{80\%}$, $W_{80\%}$ and the peak-to-plateau ratio). The absolute differences in $W_{80\%}$ (figure 3(b)) ranged from -0.1 to 0.4 mm for GATE and RayStation, and a maximum difference of -0.7 mm was observed for Eclipse. RayStation overestimated $W_{80\%}$ for most energies whilst Eclipse underestimated $W_{80\%}$ for energies higher than 150 MeV. Regarding the AUC (figure 3(c)), the absolute and mean differences across all nominal energies were within 0.4% and 0.1%, respectively, for all algorithms. GATE presented the largest standard deviation amongst all models, while Eclipse overestimated the area under the curve for most energies. Differences in dose at 2 cm depth and at the peak (figure 3(d) and (e), respectively) between the GATE model and the measured IDDs were within 2%, with a tendency to underestimate the dose at 2 cm depth and overestimate the dose at the peak. Differences in dose at 2 cm depth were within approximately 0.5% for RayStation and ranged from -2% to 0.8% for Eclipse, showing an energy dependence. For most energies, the differences in peak dose were less than 1% for both Eclipse and RayStation.

Table 2. Benchmarking data of the dose calculation models against basic experimental data: IDDs, spot profiles with and without RSs and box fields in water with varying range, width, with and without range shifting devices.

Diff. = algorithm−measured			Mean ± standard deviation [range]		
			GATE	Eclipse	Raystation
Integral depth-dose curves	Range at 80% dose (mm)		0.0 ± 0.2 [−0.3 to 0.4]	0.0 ± 0.1 [−0.1 to 0.1]	0.0 ± 0.0 [−0.1 to 0.1]
	Width at 80% dose (mm)		0.1 ± 0.1 [−0.1 to 0.3]	−0.2 ± 0.3 [−0.7 to 0.1]	0.1 ± 0.1 [−0.1 to 0.4]
	Area under curve (%)		0.0 ± 0.1 [−0.3 to 0.4]	0.1 ± 0.0 [0.0 to 0.2]	0.0 ± 0.0 [−0.1 to 0.1]
	Dose at 2 cm (%)		−1.0 ± 0.3 [−1.7 to −0.3]	−0.8 ± 0.9 [−2.0 to 0.8]	−0.2 ± 0.3 [−0.6 to 0.4]
	Dose at peak (%)		1.1 ± 0.6 [−0.2 to 2.0]	−0.3 ± 0.3 [−0.7–0.3]	−0.9 ± 1.2 [−6.0 to 0.4]
Spot profiles	Isocentre—X (mm)	No RS	0.0 ± 0.0 [−0.1 to 0.1]	0.0 ± 0.0 [−0.1 to 0.1]	0.0 ± 0.1 [−0.1 to 0.1]
		RS = 2 cm	0.0 ± 0.1 [−0.4 to 0.1]	0.0 ± 0.0 [−0.2 to 0.0]	0.0 ± 0.1 [−0.3 to 0.1]
		RS = 3 cm	−0.1 ± 0.2 [−0.7 to 0.1]	0.0 ± 0.1 [−0.2 to 0.0]	0.1 ± 0.1 [−0.2 to 0.1]
		RS = 5 cm	−0.1 ± 0.2 [−0.7 to 0.1]	−0.1 ± 0.1 [−0.2 to 0.0]	−0.1 ± 0.1 [−0.3 to 0.1]
	Isocentre—Y (mm)	No RS	0.0 ± 0.1 [−0.2 to 0.1]	0.0 ± 0.0 [−0.1 to 0.1]	0.0 ± 0.1 [−0.1 to 0.1]
		RS = 2 cm	−0.1 ± 0.1 [−0.6 to 0.1]	0.0 ± 0.0 [−0.1 to 0.0]	−0.1 ± 0.1 [−0.3 to 0.1]
		RS = 3 cm	−0.1 ± 0.2 [−0.8 to 0.1]	0.0 ± 0.1 [−0.2 to 0.0]	−0.1 ± 0.1 [−0.3 to 0.1]
		RS = 5 cm	−0.1 ± 0.2 [−0.8 to 0.3]	−0.1 ± 0.1 [−0.3 to 0.1]	−0.1 ± 0.1 [−0.3 to 0.0]
10 × 10 × 10 cm^3 box fields (without RSs)	Integral depth−dose (%)	$R_{80\%} = 15$ cm	−0.4 ± 0.9 [−1.6 to 1.2]	−0.5 ± 1.5 [−1.8 to 4.0]	−0.2 ± 0.8 [−2.7 to 1.0]
		$R_{80\%} = 20$ cm	0.0 ± 1.2 [−1.7 to 2.4]	−1.8 ± 0.9 [−3.1 to −0.3]	−0.3 ± 0.4 [−0.9 to 0.6]
		$R_{80\%} = 30$ cm	−0.2 ± 0.6 [−1.1 to 0.9]	−1.9 ± 1.9 [−6.0 to 1.5]	−0.4 ± 1.9 [−6.0 to 2.2]
	Lateral profile (%)	$R_{80\%} = 15$ cm	−0.1 ± 0.3 [−0.9 to 0.4]	−0.2 ± 1.2 [−2.0 to 1.9]	−0.3 ± 0.4 [−1.5 to 0.5]
		$R_{80\%} = 20$ cm	−0.4 ± 0.6 [−1.5 to 0.7]	−0.0 ± 1.9 [−3.3 to 2.4]	−0.4 ± 0.7 [−2.0 to 0.9]
		$R_{80\%} = 30$ cm	−0.5 ± 0.9 [−2.0 to 1.0]	−0.2 ± 1.6 [−2.9 to 2.0]	−0.1 ± 0.6 [−1.0 to 0.9]
5 × 5 × 5 cm^3 box fields $R_{80\%} = 7$ cm (with RSs)	Integral depth−dose (%)	No RS	−0.4 ± 1.2 [−2.0 to 2.5]	0.1 ± 1.6 [−1.4 to 5.4]	−0.7 ± 0.4 [−1.7 to 0.0]
		RS = 2 cm	−0.8 ± 1.1 [−2.2 to 1.9]	−0.3 ± 2.2 [−3.4 to 4.9]	−0.9 ± 0.7 [−2.3–0.6]
		RS = 3 cm	−1.4 ± 0.9 [−2.5 to 0.3]	−0.3 ± 2.1 [−3.0 to 4.2]	−1.5 ± 0.7 [−3.8 to −0.7]
		RS = 5 cm	−1.5 ± 0.7 [−2.9 to −0.3]	−1.3 ± 2.5 [−3.3 to 5.0]	−1.7 ± 0.7 [−3.4 to −0.7]
	Lateral profile (%)	No RS	−0.2 ± 0.4 [−1.6 to 0.4]	−0.9 ± 0.8 [−1.2 to 1.3]	0.6 ± 0.7 [−0.3 to 2.1]
		RS = 2 cm	−0.4 ± 0.5 [−1.6 to 0.6]	−0.1 ± 0.8 [−1.7 to 0.9]	−0.5 ± 0.5 [−1.5 to 0.3]
		RS = 3 cm	−0.3 ± 0.5 [−1.0 to 0.8]	−0.4 ± 0.8 [−1.5 to 0.8]	−0.1 ± 0.5 [−0.9 to 1.1]
		RS = 5 cm	0.3 ± 0.7 [−1.1 to 1.7]	−0.4 ± 0.7 [−1.4 to 0.5]	0.0 ± 0.6 [−1.0 to 1.7]

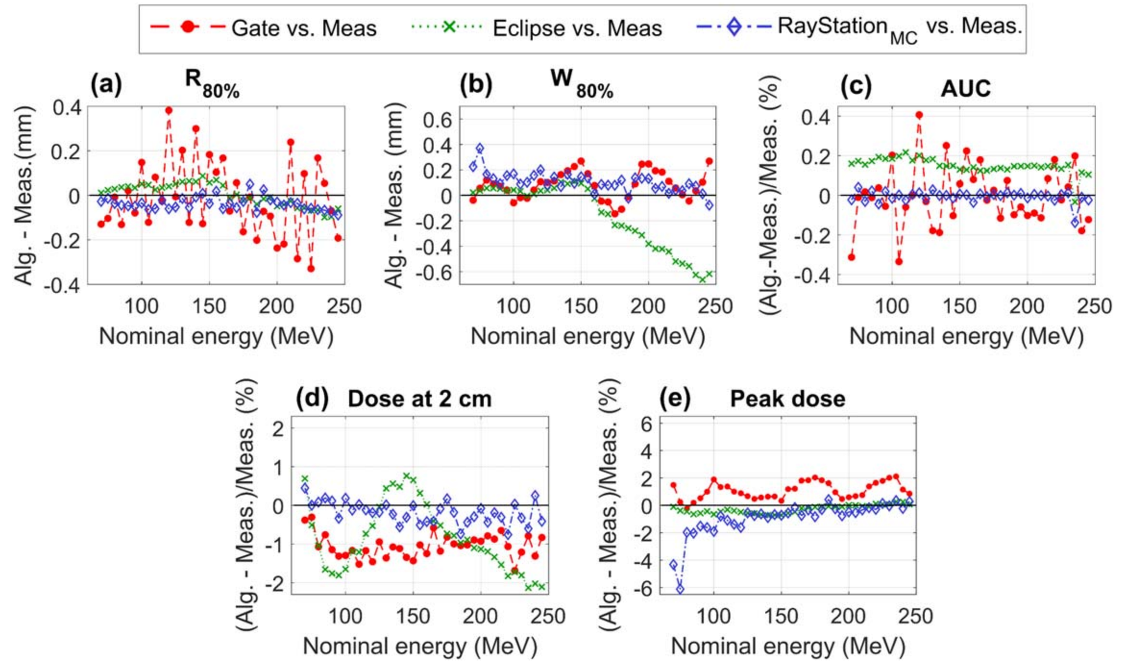


Figure 3. Comparison between the modelled and the measured IDDs in terms of $R_{80\%}$ (a), $W_{80\%}$ (b), AUC (c), dose at 2 cm depth (d) and dose at the peak (e).

RayStation underestimated the peak dose by up to 6% for the lower energies, likely because the modelled IDDs could not be exported from the system with a resolution finer than 1 mm. In comparison, the resolution of the modelled IDDs in GATE was 0.1 mm; in Eclipse it varied with depth and energy, however, it was approximately 0.1 mm in the peak region for the lower energies.

The mean differences in spot size across all nominal energies, between the measured commissioning data and the three beam models, without any RSs and for the 5 cm RS, are shown in figure 4(a) and (b), respectively, as a function of distance from the isocentre (*x* direction only). Figure 4 also shows the absolute differences in spot size at the isocentre (*x* direction) as a function of nominal energy for the three algorithms against measurements, without any RSs (c) and using the RS = 5 cm (d). The mean differences for the case without RSs were within 0.1 mm for all models (maximum absolute differences of 0.3 mm), with Eclipse having the smallest standard deviation (maximum of 0.04 mm) and RayStation the largest (maximum of 0.14 mm), considering all depths. In GATE, the highest differences occurred for the extreme energy values (70 MeV and 245 MeV), both with and without RS, likely due to a poorer fit of the parametrisation in this region. The use of RS was generally associated with larger errors in spot size. Mean differences were within 0.4 mm for all models and generally smaller mean differences were observed for the RS = 2 cm and RS = 3 cm options. Maximum absolute differences in spot size were 0.8 mm for the TPSs and 1.7 mm for GATE when the RS = 5 cm was included, and these were typically found for the measuring planes furthest from the source. For the RS = 5 cm case, maximum differences of 0.7 mm were found in GATE for the lowest energy at the isocentre, which is equivalent to approximately 3.5% of the spot size (20 mm). Differences were within 0.2 mm for energies above 120 MeV, for all RS options. For Eclipse and RayStation, all differences were within 0.2 mm and 0.4 mm, respectively, independently of RS thickness. There was a tendency for all algorithms to underestimate slightly the spot size for energies below 120 MeV, for the three RS options.

The 10 × 10 × 10 cm^3 plans with 15, 20 and 30 cm range calculated using the three models agreed well with measured data. IDDs and lateral profiles for the 20 cm range plan are presented in figures 5 (a) and (b), respectively. In general, for the IDDs measured on these box fields, GATE underestimated the dose in the build-up region and overestimated the dose in the SOBP by up to 2%. Eclipse tended to underestimate the dose, with the largest differences found for the plan with 30 cm range (up to 3.5% in the SOBP) and presented a flatter SOBP for all cases, unlike the trend seen in measurements. There was no trend for RayStation. If differences above 3.5% occurred, these were typically in the fall-off region. For the lateral profiles, differences in the tail region were the largest for Eclipse, which underestimated the dose for all plans. IDDs for the 5 × 5 × 5 cm^3 plan with RS = 5 cm (figure 5(c)) were within approximately 3% for GATE and RayStation. Similar differences were reported in other studies (Rahman *et al* 2020). Generally, GATE and RayStation underestimated the dose in comparison to measurements. Differences were within 5% for Eclipse, which overestimated the dose in the

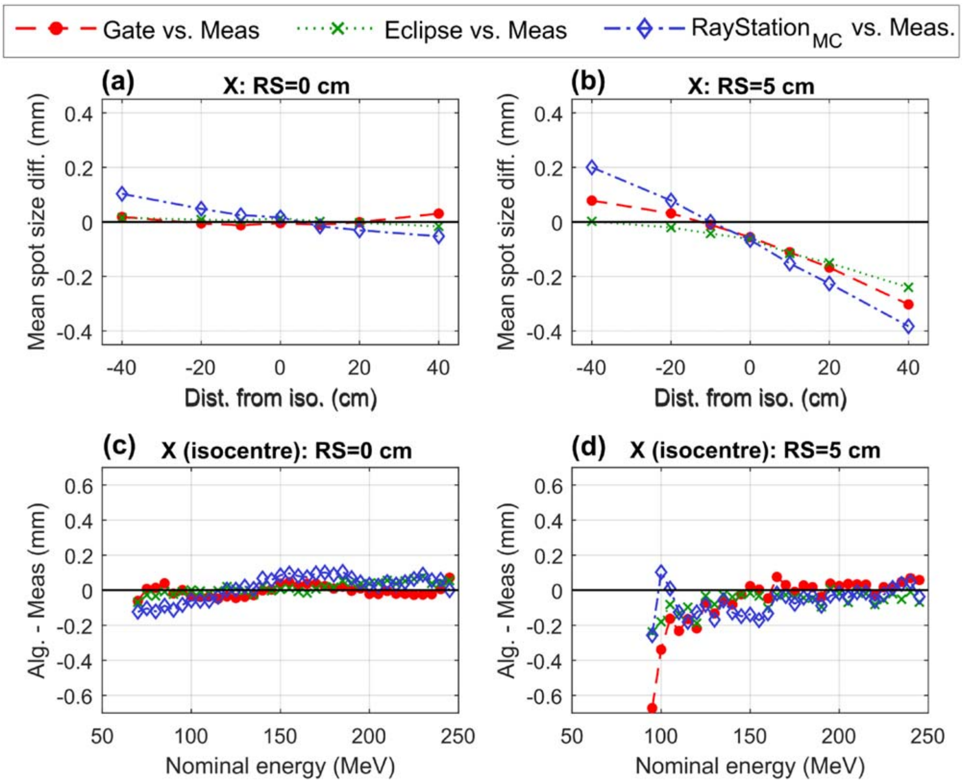


Figure 4. Comparison between the modelled and the measured spot profiles. Average spot size difference across all energies, for the seven measuring depths, for the *x* profiles, without the presence of RSs (a) and considering RS = 5 cm (b); difference between the modelled and the measured spot sizes, for all clinical energies, at the isocentre plane, for spot profiles without RSs (c) and with the RS = 5 cm (d).

build-up and underestimated the dose in the SOBP. Overall, slightly smaller differences in IDDs were achieved for plans with the 2 and 3 cm RSs. For the corresponding lateral profiles (figure 5(d)), all algorithms presented point by point differences within 1.7% for all RS options and Eclipse showed a better agreement with measurements in the tail region in comparison to the results for the plans without RSs.

3.2. Demonstration of the use of MC to complement commissioning measurements

The following sections aim to demonstrate the potential of a benchmarked MC beam model to support the different stages of the validation of the TPS in a new proton therapy facility. The outputs of Eclipse and RayStation were compared against measurements and GATE for an extensive set of plans for different scenarios: (1) optimisation of TPS calculation parameters, (2) dose assessment in homogeneous fields and (3) dose assessment in non-homogeneous clinical fields.

3.2.1. Scenario 1: optimisation of calculation parameters

Ten SOBP plans were calculated in Eclipse using a lateral cut-off (σ_{Ecl}) of 2, 3 or 4. The point dose outputs were compared to either measurements or GATE simulations. Let $\mathrm{DD}_{\mathrm{TPS\ versus\ Meas.}}$ be the dose differences obtained when comparing the TPSs with experimental data and $\mathrm{DD}_{\mathrm{TPS\ versus\ MC}}$ the dose differences between the TPSs and GATE. There was a strong correlation between $\mathrm{DD}_{\mathrm{TPS\ versus\ Meas.}}$ and $\mathrm{DD}_{\mathrm{TPS\ versus\ MC}}$ for the three σ_{Ecl} options ($\rho = 0.93$, Pearson's correlation coefficient). The mean percentage difference between Eclipse and experimental measurements was $-16.6 \pm 1.4\%$, $-2.8 \pm 0.9\%$ and $-1.5 \pm 0.9\%$ for $\sigma_{\mathrm{Ecl}} = 2$, 3 and 4, respectively. The corresponding differences against GATE were $-16.6 \pm 1.7\%$, $-2.8 \pm 1.2\%$ and $-1.5 \pm 1.3\%$, indicating very similar trends. The smallest differences between Eclipse and measurements were for $\sigma_{\mathrm{Ecl}} = 4$, and this conclusion could be derived by comparing Eclipse to GATE. The value of $\sigma_{\mathrm{Ecl}} = 4$ was used for all subsequent Eclipse dose calculations.

The ten SOBP plans were also recalculated in Eclipse for a grid resolution of $1 \times 1 \times 1$, $2 \times 2 \times 2$ and $3 \times 3 \times 3$ mm^3. The dose grid resolution had an impact only on fields with small $W_{80\%}$, where finer resolutions improved the dose calculation accuracy—in these cases, the excess error for using $3 \times 3 \times 3$ mm^3 was 0.6%. The same trend was observed when comparing Eclipse dose outputs directly to GATE results. The Pearson's correlation coefficient between $\mathrm{DD}_{\mathrm{TPS\ versus\ Meas.}}$ and $\mathrm{DD}_{\mathrm{TPS\ versus\ MC}}$ was 0.95. For the seven PSQA cases tested,

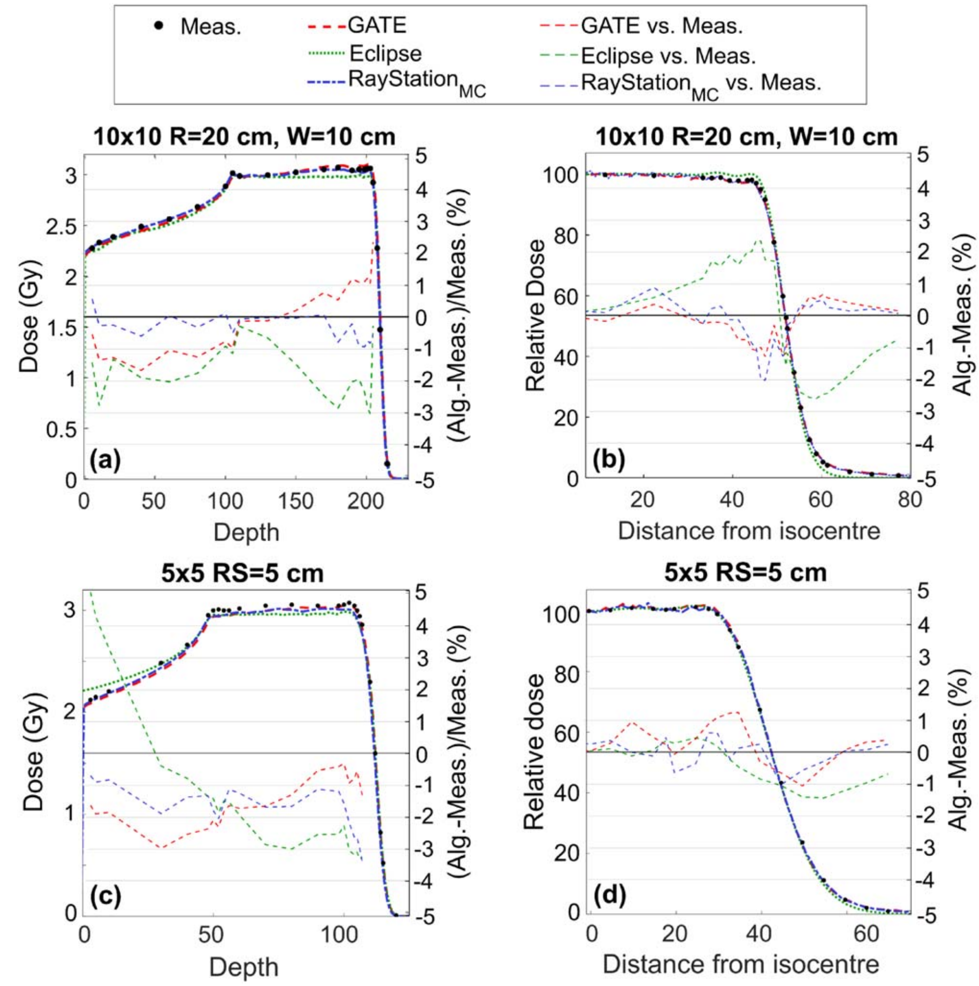


Figure 5. IDDs and lateral profiles for the 10 × 10 cm^2 SOBP plan with 20 cm range (10 cm width), without RS (subfigures (a) and (b)) and for the 5 × 5 × 5 cm^3 SOBP plan with RS = 5 cm (subfigures (c) and (d)). Differences between dose calculation algorithms and measurements are represented by the dashed lines.

the mean absolute percentage differences between doses calculated in Eclipse and measurements were 1.3 ±1.0%, 1.4 ± 1.2% and 1.8 ± 2.6%, while between Eclipse and GATE the differences were 1.4 ± 1.0%, 1.6 ± 1.2 and 1.9 ± 2.5%. The two datasets, $\mathrm{DD}_{\mathrm{TPS\ versus\ Meas.}}$ and $\mathrm{DD}_{\mathrm{TPS\ versus\ MC}}$, were strongly correlated as well for all grid resolutions ($\rho = 0.84$ Pearson's correlation coefficient). The standard deviation of the differences increased with increasing grid spacing—for example, an excess dose difference of 10% was observed for a field with large dose inhomogeneities when using a dose grid of 3 × 3 × 3 mm^3. In summary, while dose outputs extracted from the 1 × 1 × 1 mm^3 agreed best with measurement and GATE for both homogeneous and heterogeneous fields, using such a fine grid was more important for heterogeneous fields, where the errors in the positioning of the point dose are larger. A resolution of 1 × 1 × 1 mm^3 was applied to the rest of the plans calculated in this work using Eclipse and RayStation.

3.2.2. Scenario 2: Homogeneous fields

Figure 6 (a) shows normalised dose values at 2 cm depth for a 100 MeV monoenergetic layer of field sizes ranging from 3 × 3 cm^2 to 15 × 15 cm^2, obtained through measurements, GATE, Eclipse and RayStation. All dose values were normalised to the reference field size of 10 × 10 cm^2. The measured dose generally increased with increasing field size and a similar trend was observed for GATE. Surprisingly, the dose for the largest field size (15 × 15 cm^2) was 0.2% lower than for the 12 × 12 cm^2 field. However, this difference was within the uncertainty limits of the measurements and all calculation algorithms. For both Eclipse and RayStation, the dose was constant for field sizes larger than 4 × 4 cm^2. Figure 6(b) shows the percentage difference for the three dose calculations algorithms in comparison to measurements and figure 6(c) shows the percentage differences in dose for Eclipse and RayStation against GATE. For GATE, maximum differences of approximately 1.4% against measurements were observed for the smaller field sizes of 3 × 3 cm^2 and 4 × 4 cm^2. The differences for Eclipse and RayStation against measurements were comparable with those detected with comparisons against GATE—

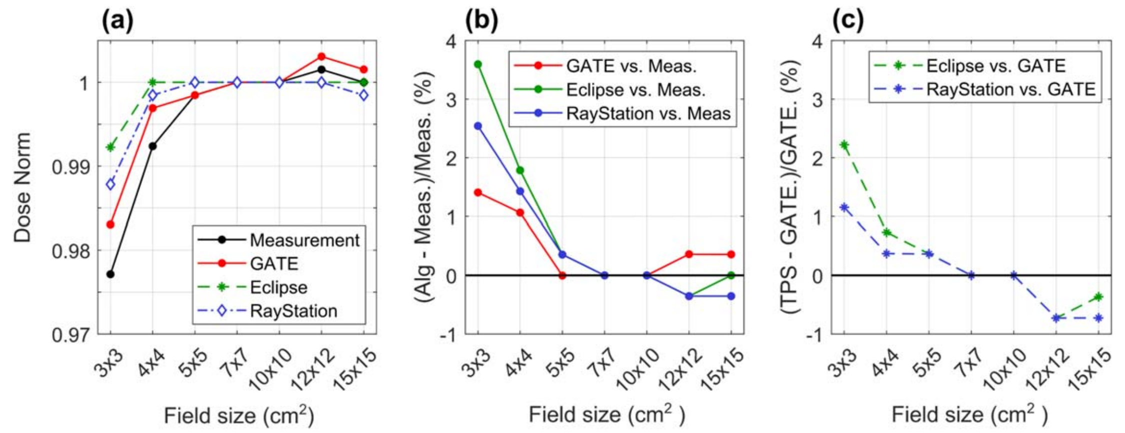


Figure 6. Normalised dose at 2 cm depth obtained through measurements and calculated using GATE, Eclipse and RayStation, for a 100 MeV monoenergetic layer, as a function of the field size (a); percentage dose difference for GATE, Eclipse and RayStation against measurements (b); percentage dose difference for Eclipse and RayStation against GATE (v).

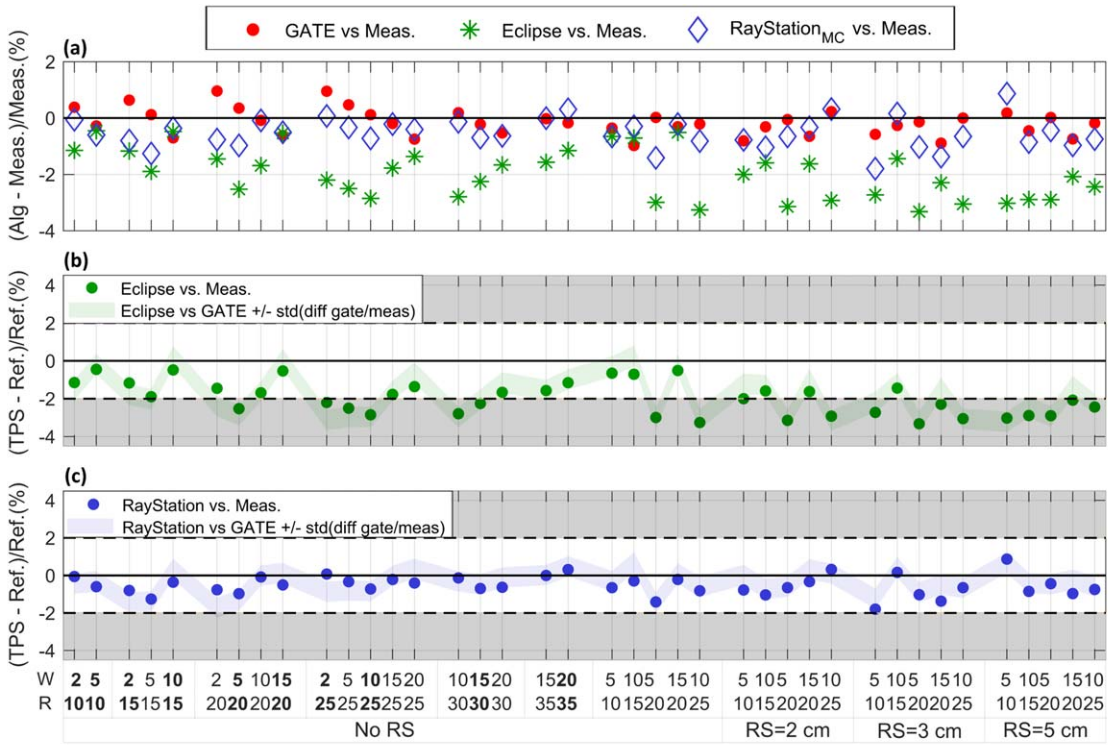


Figure 7. Percentage differences in dose at the centre of 39 SOBP plans with varying SOBP range, R, (10 cm to 35 cm) and width, W, (2 cm to 20 cm), between measurements and the three dose calculations algorithms (a); point dose differences between the TPSs and measurements and the correspondent difference region estimated by GATE, for Eclipse (b) and RayStation (c). The GATE estimated intervals were obtained from the TPSs difference against GATE plus or minus the standard deviation of the GATE differences against measurements for the 10 subsets of fields. Plans marked in bold represent the selected sample of 10 SOBP plans.

i.e. increased dose differences with decreasing field size, following very similar trends. A maximum difference of 3.5% and 2.5% was obtained for Eclipse and RayStation, respectively, when comparing with measurements. These differences were underestimated by 1.4% when comparing against GATE. In this experiment with monoenergetic layer fields, the Pearson's correlation coefficient between $\mathrm{DD}_{\mathrm{TPS\ versus\ Meas.}}$ and $\mathrm{DD}_{\mathrm{TPS\ versus\ MC}}$ was also strong ($\rho = 0.97$ for Eclipse and $\rho = 0.90$ for RayStation).

Figure 7(a) presents the percentage differences in the output dose for the three dose calculation algorithms against measurements for the total of 39 SOBP plans. The corresponding mean differences were $-0.1 \pm 0.5\%$, $-2.0 \pm 0.9\%$ and $-0.5 \pm 0.5\%$, for GATE, Eclipse and RayStation respectively. GATE presented larger differences for SOBP with smaller *W* of 2 cm and Eclipse underestimated the dose for most fields. Figure 7 also shows the dose differences at the centre of the 39 SOBPs for Eclipse (b) and RayStation (c) versus measurements

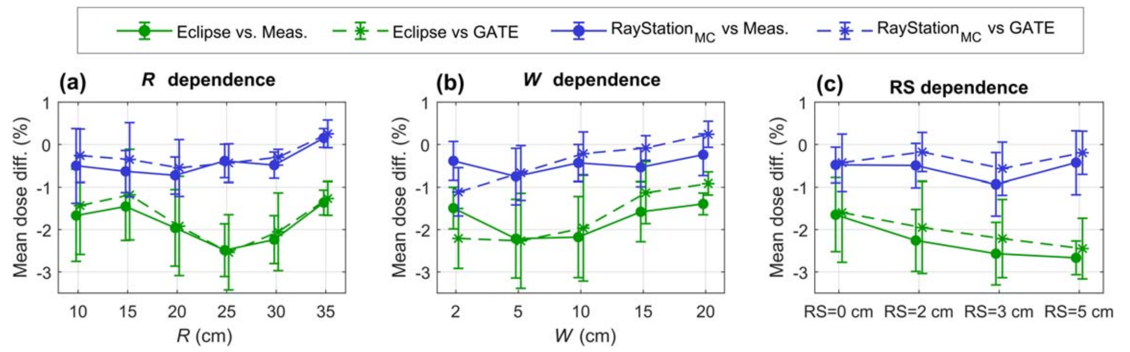


Figure 8. Average difference in dose between the TPS and measurements (solid line) and the TPS and GATE (dashed line) across all SOBP considered, as a function of R (a), W (b) and RS option (c).

(solid dots) or GATE (shaded area). The shaded area was created by applying a tolerance to the dose differences ($DD_{\text{TPS versus MC}}$), which was defined as the standard deviation of the differences between GATE and measurements (±0.54%) for a subset of 10 fields (marked in bold in the axis of the figure). The Pearson's correlation coefficient between $DD_{\text{TPS versus Meas.}}$ and $DD_{\text{TPS versus MC}}$ was 0.92 for Eclipse and 0.68 for RayStation, indicating that GATE could better identify dose differences between Eclipse and measurements.

The coloured shaded regions in figures 7(b) and (c) represent the difference between the TPSs and GATE plus a tolerance to account for the fact that GATE itself presents a difference against measurements. Ideally, the solid dots curves would fall within the confidence region of GATE (shaded coloured region). Out of the 29 points evaluated (i.e. after excluding the 10 plans used to define the tolerance), 21 points (72%) were within the GATE prediction shaded area and 8 points (28%) were outside, for both Eclipse and RayStation. Generally, the points were close to fall within the shaded area and the maximum difference between the solid dots and the border of the shaded area was approximately 0.4% for both Eclipse and RayStation. The grey areas in subfigures (b) and (c) are the regions outside the acceptance criteria established for these plans (maximum 2% difference). If the shaded region overlaps the grey region, based on comparison with GATE, there is a likelihood that the difference will fall within the non-acceptance region. In total, 19 points were outside the 2% acceptance criteria when comparing Eclipse to measurements. According to the comparison with GATE, 24 points were predicted to be outside the acceptance criteria and 17 of these points (out of 19) were correctly predicted. No points were outside the acceptance criteria when comparing RayStation to measurements, while one point had a small likelihood of being outside when comparing RayStation directly to GATE. These results show that homogenous fields simulated in a properly commissioned MC system can be used to predict TPS deviations from measurements for validation purposes, since there were no cases for which point dose differences were within 2% when compared to GATE but outside tolerance when compared to experiments. This would prevent the need for measuring the entire range of fields, and rather a focus could be made on the situations of predicted failure, reducing the amount of in-person time required for physical measurements.

Figure 8 shows the mean dose differences for Eclipse and RayStation, considering the 39 SOBP plans, against measurements (solid line) or GATE (dashed line), as a function of R, W and RS option, which allow to identify trends and limitations of the TPS in the dose calculation of different field types (deep/shallow, wide/thin, with or without RS). For instance, Eclipse presented the largest differences for plans with 25 cm R and there was a trend for differences to increase with increasing RS thickness—both limitations could be identified through comparisons with GATE alone. The largest disagreement between $DD_{\text{TPS versus Meas.}}$ and $DD_{\text{TPS versus MC}}$ was for $W = 2$ cm, in agreement with the results in figure 7 (a) where it was shown that GATE presented larger differences for $W = 2$ cm.

3.2.3. Scenario 3: non-homogeneous fields

Figure 9(a) shows the percentage differences in dose measured in up to three points per field for 11 PSQA plans (a total of 72 points) for GATE, Eclipse and RayStation against measurements. The mean differences across the 72 points were −0.7±1.2%, 0.4±1.9% and −0.3±1.0% for GATE, Eclipse and RayStation, respectively. The absolute maximum differences against measurements found were 2.9% for both GATE and RayStation and 4.5% for Eclipse. Both GATE and RayStation tended to underestimate the dose for plans with the RS = 5 cm, in comparison to plans without or with thinner RSs. In fact, the two algorithms presented a similar trend for differences against measurements across the entire dataset. Eclipse tended to overestimate the dose for plans without any RSs and underestimate the dose for shallow fields with the RS of 5 cm (Breast case). Figures 9(b) and (c) show the differences for Eclipse and RayStation, respectively, when comparing against measurements (solid

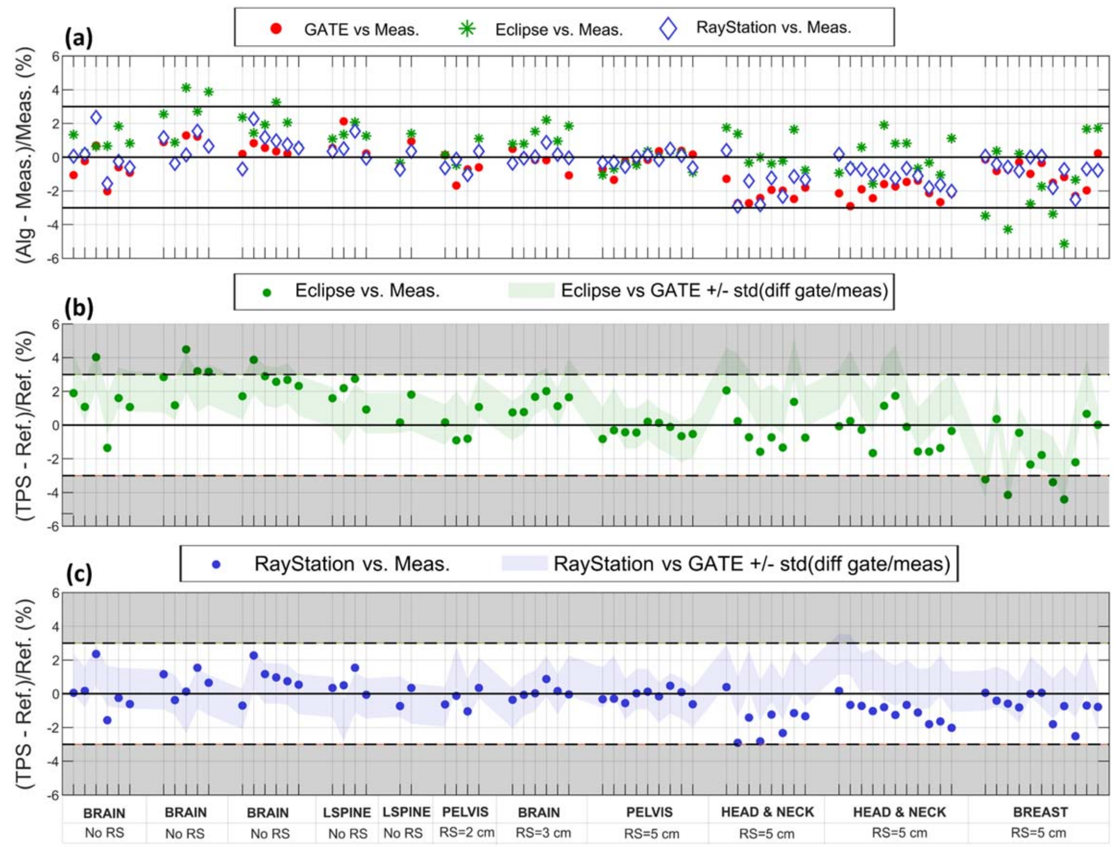


Figure 9. PSQA point doses differences for the three dose calculation algorithms against measurements (a); differences between the TPS and measurements and the correspondent difference region estimated by GATE, for Eclipse (b) and RayStation (c).

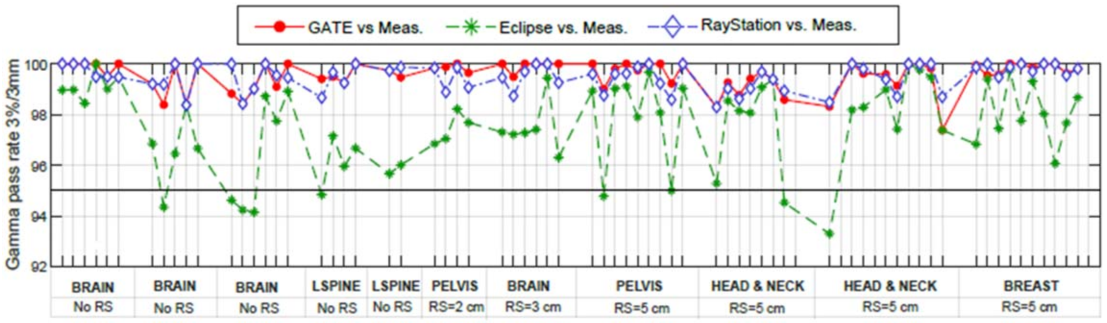


Figure 10. Local gamma pass rates (3%/3 mm, with 5% lower threshold cut) for GATE, Eclipse and RayStation against the measured planes. The data points are not correlated, and each point corresponds to an independent result within a treatment field.

dots) and GATE (shaded area). The standard deviation of the differences between GATE and measurements was ±1.2% and this value was applied as the tolerance interval when comparing TPSs dose outputs directly to GATE, similarly to what was done in the case of homogenous fields. Out of the 72 points, 43 points (60%) were within the GATE prediction shaded area and 29 points (40%) were outside, for both Eclipse and RayStation. Out of the 29 points that were outside the GATE predicted area, 23 (~80% of the points) corresponded to fields containing the RS = 5 cm, for which GATE presented larger discrepancies in comparison to measurements, whilst still within the established acceptance interval (maximum of 3% difference).

The 3%/3 mm gamma pass rate results for GATE, Eclipse and RayStation are presented in figure 10. The gamma pass rates for GATE and RayStation calculated planes in comparison to measured planes were all above 97% and 98%, respectively. Both the point dose differences (figure 9 (a)) and the gamma pass rate results for the two algorithms followed a similar trend, and this is most likely due to both algorithms being MC-based. For most plans, the gamma pass rates fluctuated between 95% and 100% for Eclipse, with 8 out of 72 points

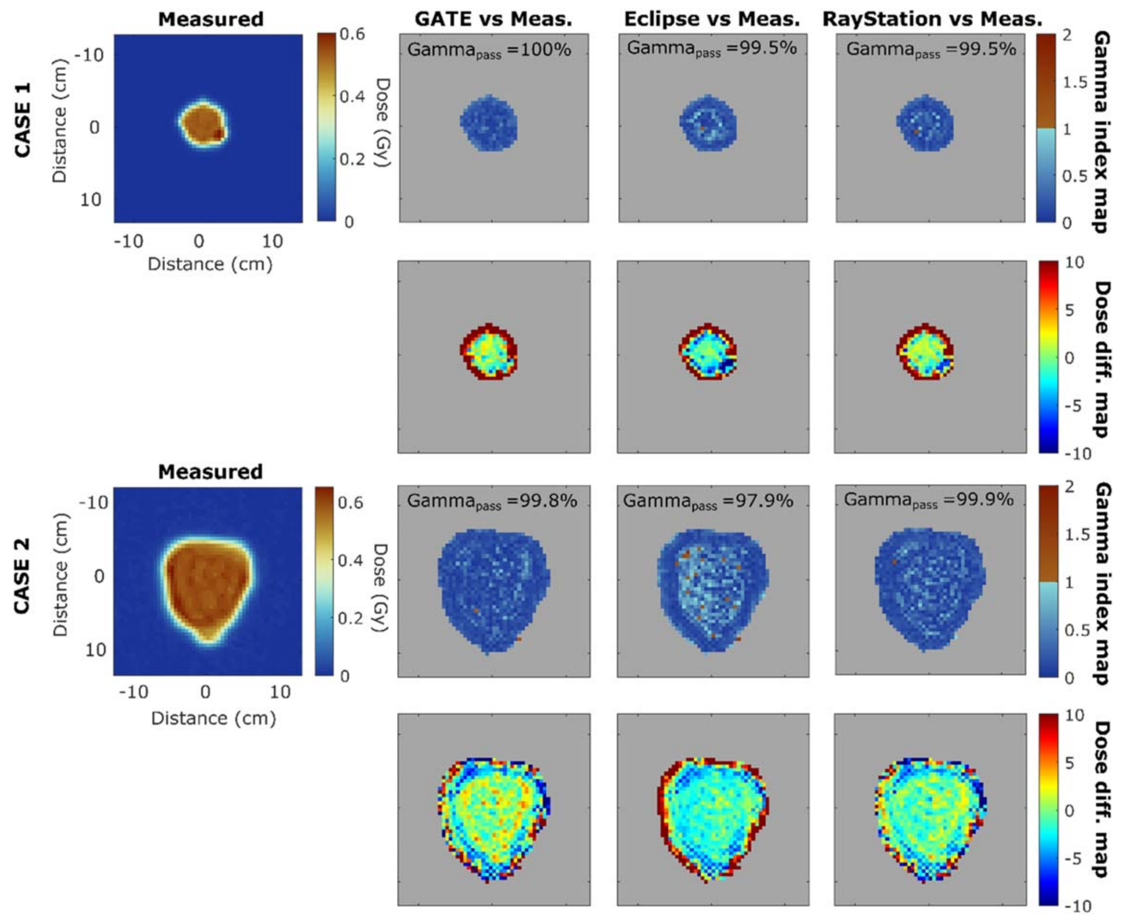


Figure 11. Examples of local 3%/3 mm gamma index and local dose difference maps between the measured planes and the three dose calculation algorithms, for two fields; case 1 is a brain case without RS and case 2 is a pelvis case with the 5 cm RS.

presenting gamma rates slightly below the established passing criteria. Although all algorithms showed different trends for cases with or without RSs for the point dose evaluation (figure 9 (a)), the same trends were not directly translated into the gamma pass rate results with the chosen specifications.

The analysis of the 2D dose planes indicated that overall, GATE and RayStation tended to underestimate slightly the dose whereas Eclipse tended to overestimate it. Examples of 3%/3 mm gamma index and dose difference ((meas.-alg.)/alg.) maps can be found in figure 11 for a clinical brain plan without RS and for a pelvis plan with RS = 5 cm.

4. Discussion

In this study we have demonstrated the potential of Monte Carlo to support the commissioning of the treatment planning system of a new proton beam therapy machine. A MC model may be developed early in the commissioning process using the same beam data required to commission a new PBS system. The MC and TPS models should first be benchmarked against commissioning data and comprehensive measurements on a small number of representative homogeneous fields to verify the accuracy of the implementations. Then, by evaluating the dose calculation algorithms on an extensive set of homogenous and non-homogenous plans, we have shown that MC may be used as an independent dose calculation tool to complement (and potentially reduce) the number of measurements during the TPS dose validation. MC has the potential role to identify the parameter space in which the TPS is expected to deviate from measurements and so focus in-person measurement efforts on these cases for best use of commissioning time. Furthermore, it can help understand the limitations and outputs from the TPSs, as well as in inform the optimisation of the clinical dose calculation algorithms. This potential was demonstrated for different dose calculation engines available in two commercial TPS systems. To the best of our knowledge this is the first study to focus on the use of MC to support the dose validation and verification steps of the commissioning of a treatment planning system.

The first part of the work demonstrated a semi-automated process to develop a proton beam model in GATE and proposed a set of detailed measurements to benchmark its performance. This methodology is generalisable and could be applied to model beam data from other PBS-PT centres with similar technology. The beam modelling methodology applied in this work was based on that of Yeom *et al* (2020). This consisted of an iterative optimisation of the energy and optical properties of the beam and the final model was a parametrisation of the optimal beam parameters as a function of nominal energy. To model the optical parameters in GATE, the best initialisation parameters were first roughly estimated, which accelerated the convergence of the optimisation. The IDDs were calibrated using the area under the curve, thus avoiding the normalisation to be performed at a single point along the IDD, similar to Aitkenhead *et al* (2020). One limitation of this modelling approach is the parametrisation itself, as the error between the optimal values for the beam parameters and the fitted values can be considerable (up to 10%), particularly for the divergence and the energy spread. Similar findings were observed by Grevillot *et al* (2011) and Aitkenhead *et al* (2020) for the energy spread parameter, although exact error values were not reported in the publications.

Maximum differences of 0.3 mm were found when comparing spot sizes without RSs obtained with GATE, Eclipse and RayStation to measurements, for all measuring depths, and maximum differences ranging from 0.15 to 0.4 mm were reported in the literature (Grevillot *et al* 2011, Rahman *et al* 2020, Saini *et al* 2017, Yeom *et al* 2020). An underestimation of the spot sizes of profiles with RSs was observed for all algorithms, for energies below 120 MeV. The measured profiles were noisier for the lower energies, therefore, there is a larger uncertainty associated to these measured spot sizes. Differences in spot size against the air profiles obtained during commissioning were slightly lower for Eclipse in comparison to GATE and RayStation, which may be related to the way the different systems model the RSs. Eclipse system uses as input the measured spot profiles with and without RSs in the beam modelling process. However, in both GATE and RayStation the commissioning data of spot profiles with RSs were not used, and only the material of the RSs was modelled. In RayStation, the vendor optimised material was provided within the material options. In GATE, the density of the RS material was tuned to match the measured WET, however, this could be further improved to better model the true scattering properties of the material, perhaps by tuning the exact chemical composition and I-value. In our MC beam modelling process, we optimised optical parameters to match only the experimental data without RS. Improvements could be achieved by, for example, finding the best optical parameters that match spot profiles both with and without RS, or to generate an independently optimised model for each RS separately (Fracchiolla *et al* 2015, Winterhalter *et al* 2020a).

The differences in $R_{80\%}$ between the measured and the modelled Bragg peaks were within 0.4 mm for all algorithms and maximum differences of 0.6 mm and 1 mm have been reported in the literature (Grevillot *et al* 2011, Saini *et al* 2017). These small differences were translated into errors in range of the SOBP plans within the calculated dose grid resolution. The overall shape of the IDDs and the way the absolute dose calibration is implemented may reflect on the performance of the models. In GATE, the IDDs were calibrated to the area under the curve, therefore there is a balance between the agreement in build-up and peak regions, when comparing to measurements. The underestimation of the dose at 2 cm depth in the IDDs was reflected in the lower dose in the build-up region of the SOBPs and the overestimation of the dose at the peak can be associated with the higher dose in the flat region of the SOBPs. Additionally, the overestimation of the dose in the peaks of the IDDS is reflected in the dose outputs in the centre of SOBP (figure 7). This overestimation was larger for SOBP fields with smaller width, where a greater proportion of the dose is coming from the peak region, and decreased with increasing width, where there is a larger contribution from the build-up regions of the individual beams. Despite a good agreement against measurements of the IDDs peak dose in Eclipse, a flat high dose region in SOBP plans was observed, unlike the pattern of the measurements or GATE and RayStation. Furthermore, it underestimated the dose in the centre of SOBPs by up to 4%. This can be associated with the fact that no correction factor was applied from these box-field results, which is a possible dose calculation refinement in Eclipse. We opted against applying this correction to our Eclipse beam model since the dose underestimation found for the homogeneous box-fields did not propagate to non-homogeneous fields.

It is of utmost importance to compare the final beam models built both in the TPSs and MC against the commissioning data itself, as any discrepancies present at this stage will be reflected directly on more complex homogeneous and non-homogeneous clinical plans. For instance, the current version of RayStation does not automatically compute spot profiles in the presence of the RSs, since this data is not used to build the models. Users must perform the dose calculations of monoenergetic pencil beams in air for the full energy range and extract the corresponding spot sizes independently, and from our experience this should not be overlooked. In an earlier version of our RayStation dose model we found large differences against both measurements and GATE for non-homogeneous complex fields with RS for which were struggling to find a justification. It was upon explicitly benchmarking the air profiles with RS that we realised an error in defining the distances between the isocentre and RS tray position. This error was subtle when analysing simpler, homogeneous fields. Having a benchmarked MC when we started the commissioning process of RayStation was crucial to identify (and correct for) this error. More details on the differences in spot size in the presence of RSs pre- and post-correction of the RS position can be found in supplementary material 2.

When benchmarking the beam models, we tested their performance in seven representative SOBPs. Measuring IDDs and lateral profiles in SOBP fields is time consuming. We believe that performing these measurements in four SOBPs only would provide a good understanding of the models' performance (three SOBPs plans of different ranges and one SOBP plans with the thickest RS). Other SOBPs with different configurations could be tested based on MC. In the second part of this work, an extensive range of measurements was performed to demonstrate the potential of MC to support the TPS dose validation and evaluation, therefore, not exclusively to validate the MC and TPS models. First, it was shown that MC can help optimise TPS calculation parameters using a limited number of experimental data (10 SOBP fields). The conclusion regarding the most suitable TPS dose calculation parameters, like lateral cut-off σ_{Ecl} and grid resolution, was straightforward from MC, therefore we believe that calculation parameters can be chosen based on comparisons of dose outputs in homogeneous and non-homogeneous fields against MC only, without the need for measurements. Additionally, MC can be used to understand the TPS performance in homogeneous fields, where from a smaller number of plans measured experimentally (the same 10 SOBP plans), one could gain confidence on the performance of the TPS on a wider range of fields (in this work, we investigated 29 more plans). For our delivery system, selecting and measuring only a quarter of the total field of interest was adequate to find a GATE acceptance interval applicable to most fields. Furthermore, it was shown that MC can help inform on the impact of aspects such as field size, range and width of SOBPs and the use of RS in complex plans. Regarding the experiment which aimed to understand the dose output variation in monoenergetic layers of different field size, MC could potentially be used to explore the dose variation for other energies and other depths along the Bragg peak, although we did not investigate this in our study. Finally, MC can support the early development and streamline of PSQA processes. Such a system can allow a more efficient and thorough exploration of the TPS's performance over the full range of clinically relevant scenarios and help identify any limitations ahead of going live.

Rich experimental data helps building confidence in the dose models used clinically. We found maximum differences against experimental measurements and the three beam models to be within 3.3% in homogeneous fields, and 4.5% in non-homogeneous fields. This is in agreement with values reported in the literature (Trnková *et al* 2016, Winterhalter *et al* 2018, Aitkenhead *et al* 2020). It is important to add that although experimental measurements are considered the gold standard in dosimetry, these also have an associated uncertainty. One source of uncertainties are the detectors, which are susceptible to positioning and setup errors. Detectors may have calibration uncertainties and variations due to the operation and environmental conditions, which may be also accompanied by beam output variations. Coutrakon *et al* (2010) estimated the error in dose delivered to a water phantom by introducing multiple random beam delivery errors and calculating the root mean square of the dose variation. The authors verified that dose errors due to beam energy and spot positioning variations could be approximately 1.85%; errors due to beam spill non-uniformity, intensity regulation and finite scanning speed were below 0.5%. Although dosimetry uncertainties are recommended to be as low as possible, these could add up to approximately 2% (Arjomandy *et al* 2019).

It was demonstrated that MC tools have the potential to complement the time-consuming measurements, as long as there is an awareness and confidence about the level of uncertainty of the MC model itself. The strong correlation coefficient between $\mathrm{DD}_{\mathrm{TPS\ versus\ Meas.}}$ and $\mathrm{DD}_{\mathrm{TPS\ versus\ MC}}$ for homogeneous fields indicated that TPS dose can be confidently compared with MC to complement measurements. A tolerance was identified on a smaller number of fields (10 fields), and it was applied to understand dose differences on a wider range of scenarios (29 additional fields). By using a carefully calibrated MC system, one can study many more scenarios than those that can be measured due to time constraints and gain a deeper understanding of the TPS system and its limitations. In our work, most point dose outputs that fell outside the clinically established passing criteria when comparing to measurements, were also outside the same criteria when comparing against GATE. Having a preliminary knowledge from MC about the expected measurements may help understand and anticipate what types of fields are more likely to not meet set tolerance criteria and inform the need for additional refinement of the TPS models.

MC workflows are being increasingly used as independent dose calculation tool for PSQA processes (Aitkenhead *et al* 2020, Xu *et al* 2022), and MC can also be applied during commissioning to support the early development of PSQA protocols used at each centre. We analysed a small number of PSQA plans for a variety of anatomical sites for proof-of-concept purposes. Unlike for homogeneous plans, where one tolerance value was applied for TPS comparisons against GATE and it was applicable to most plans, the threshold established as tolerance for the PSQA results (standard deviation of differences between GATE and measurements for the 11 cases) was not adequate for all types of plans. Having a comprehensive analysis of both TPS and MC doses for a small number of cases, can give some early indication of the dose differences between dose models and measurements for different clinical sites or treatment configuration and help the clinical teams decide on the most adequate processes and criteria for PSQA. For example, our findings suggests that some plan configurations, like typical head and neck fields with RS, may require different thresholds for MC to be use as an independent dose calculation algorithm. Furthermore, once we started patient treatment in our institution, we realised that Eclipse tended to overestimate the point dose outputs for smaller field plans without RSs and underestimate the dose outputs for large shallow plans with RSs, like breast plans. These discrepancies were not expected prior to collecting measurements for a significant number of plans, but having the MC information, we could have had better insights on the PSQA procedure to adopt for these cases. Furthermore, planes not passing the established criteria of 95% for the gamma pass rate did not necessarily fail the passing criteria of 3% for the point dose measurements within the same 2D plane and vice-versa. More plans for the same anatomical site and with similar treatment configurations must be evaluated to gain better confidence and reproducibility or to identify any peculiar trends in the dose calculation outputs.

The commissioning period in very intensive and there may be a compromise in the number of measurements that would ideally be performed for the TPS dose model validation. Additionally, it may also be hard to identify when enough measurements have been done to be confident in the dose model. Measuring IDDs and lateral profiles in SOBP fields or clinical PSQA fields is extremely time consuming as the entire field must be redelivered for each measurement point. Full experimental validation for a range of different field sizes, depths, axes of measurement, RSs options, etc, would require numerous days of measurements. Due to constraints on commissioning timelines and staffing, the full set of planned measurements may not be performed, which reduces the chances of identifying any limitations in the TPS beam model ahead of going clinical, risking the clinical acceptance of a non-optimal solution. We believe that MC can help reaching the confidence level in the TPS dose model quicker. MC can help troubleshoot if any discrepancies are present in the TPS model, test if tuning TPS parameters will improve model accuracy, and overall explore more scenarios than those that can be realistically verified experimentally.

There are further benefits to having a tailored MC model once a facility is clinically operating. MC can also be used to support translational research work on applications such as linear energy transfer and relative biological effectiveness calculations (non-available in all TPSs) (Smith *et al* 2022), out-of-field dosimetry studies and assessment of radiation-induced late effects (Yeom *et al* 2020, De Saint-Hubert *et al* 2022). Additionally, it is common practice for centres with multiple gantries to commission these sequentially and acquire first all the commissioning data required to build a beam model in the TPS from a single gantry, whilst the other gantries are being installed. Ideally, the beam properties would match exactly across all gantries, however, in practice, this will not be the case and there will always be some discrepancies in spot sizes, outputs, range, etc. An established process in-house for automated MC modelling can also facilitate future work evaluating the impact of differences between gantries and a refined beam model which provides a more representative match to all gantries could be created.

In summary, in this work, we have demonstrated that an adequately benchmarked MC model, developed early in the commissioning of a new PBT facility, can support the commissioning of the TPS on different applications, including optimisation of TPS calculation parameters, understanding of the dose calculation limitations and early development of PSQA protocols. However, regardless of the advantages that MC brings in both the shorter- and longer-terms, building a MC beam model may not be viewed as a priority during the busy commissioning period, particularly due to lack of in-house MC expertise. However, MC methods for beam modelling are becoming increasingly available and shared and commercial products of MC tools as independent dose calculations are becoming available (Fuchs *et al* 2021). The detailed description of the MC implementation process and evaluation of its performance and limitations on a comprehensive range of experimental data, as presented in this study, along with the need for developing tools to facilitate, advance and automate commissioning steps, will help proton centres achieve shorter commissioning periods and streamline their daily work.

5. Conclusions

In this work, we developed a MC model in GATE of the clinical beam at our institution and investigated how that MC could be used to support the extensive and time-consuming experimental measurements during the commissioning of the TPS system in a new proton therapy facility. We compared two commercial TPSs with different dose calculation engines (Eclipse PCS and RayStation MC), against experiments and GATE, for an extensive set of homogeneous plans in water and non-homogeneous PSQA fields in solid water. The three beam models were first benchmarked against experimental measurements, which verified their performance to be within clinically acceptable limits. This work demonstrates that establishing a MC system early on in the commissioning process can greatly enhance a centre’s ability to fully explore the performance and limitations of their TPS by reducing the number of time-intensive measurements that must be performed. It may also support the development of PSQA processes and acceptance criteria for different sites ahead of treatment start.

Acknowledgments

DB was supported by Fundação para a Ciência e Tecnologia (SFRH/BD/145287/2019). CV was supported by the Royal Academy of Engineering under the Research Fellowship scheme (RF\201718\17140). The work performed at NPL was supported by the National Measurements System Programmes Unit of the UK's Department for Science, Innovation and Technology. The authors would like to acknowledge the contribution of Suzanne Broadman and Valerios Politidis for their initial contribution to beam modelling and data analysis.

Data availability statement

The data cannot be made publicly available upon publication because they are owned by a third party and the terms of use prevent public distribution. The data that support the findings of this study are available upon reasonable request from the authors.

References

Agostinelli S *et al* 2003 GEANT4—a simulation toolkit *Nucl Instrum. Methods Phys.*A **506** 250–303

Aitkenhead A H, Sitch P, Richardson J C, Winterhalter C, Patel I, Mackay R I *et al* 2020 Automated Monte-Carlo re-calculation of proton therapy plans using Geant4/Gate: implementation and comparison to plan-specific quality assurance measurements *Br. J. Radiol.* **93** 20200228

Arjomandy B *et al* 2019 AAPM task group 224: comprehensive proton therapy machine quality assurance *Med. Phys.* **46** e678–705

Böhlen T T, Cerutti F, Chin M P W, Fassò A, Ferrari A, Ortega P G, Mairani A, Sala P R, Smirnov G and Vlachoudis V 2014 The FLUKA Code: developments and challenges for high energy and medical applications *Nucl. Data Sheets* **120** 211–4

Clasie B, Depauw N, Fransen M, Gom C, Panahandeh H R, Seco J, Flanz J B and Kooy H M 2012 Golden beam data for proton pencil-beam scanning *Phys. Med. Biol.* **57** 1147–58

Coutrakon G, Wang N, Miller D W and Yang Y 2010 Dose error analysis for a scanned proton beam delivery system *Phys. Med. Biol.* **55** 7081–96

de Martino F, Clemente S, Graeff C, Palma G and Cella L 2021 Dose calculation algorithms for external radiation therapy: an overview for practitioners *Appl. Sci.* **11** 6806

De Saint-Hubert M *et al* 2022 Validation of a monte carlo framework for out-of-field dose calculations in proton therapy *Front. Oncol.* **12** 1–15

Farr J B *et al* 2021 Clinical commissioning of intensity-modulated proton therapy systems: report of AAPM Task Group 185 *Med. Phys.* **48** e1–30

Foote R L *et al* 2012 The clinical case for proton beam therapy *Radiat. Oncol.* **7** 1–10

Fracchiolla F, Lorentini S, Widesott L and Schwarz M 2015 Characterization and validation of a Monte Carlo code for independent dose calculation in proton therapy treatments with pencil beam scanning *Phys. Med. Biol.* **60** 8601–19

Fuchs H, Elia A, Resch A F, Kuess P, Lühr A, Vidal M, Grevillot L and Georg D 2021 Computer-assisted beam modeling for particle therapy *Med. Phys.* **48** 841–51

Gottschalk B, Cascio E W, Daartz J and Wagner M S 2015 On the nuclear halo of a proton pencil beam stopping in water *Phys. Med. Biol.* **60** 5627–54

Grassberger C, Lomax A and Paganetti H 2015 Characterizing a proton beam scanning system for monte carlo dose calculation in patients *Phys. Med. Biol.* **60** 633–45

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** 5203–19

Grevillot L, Bertrand D, Dessy F, Freud N and Sarrut D 2012 GATE as a GEANT4-based monte carlo platform for the evaluation of proton pencil beam scanning treatment plans *Phys. Med. Biol.* **57** 4223–44

Guterres Marmitt G, Pin A, Ng Wei Siang K, Janssens G, Souris K, Cohilis M, Langendijk J A, Both S, Knopf A and Meijers A 2020 Platform for automatic patient quality assurance via Monte Carlo simulations in proton therapy *Phys. Med.* **70** 49–57

Harms J, Chang C W, Zhang R and Lin L 2020 Nuclear halo measurements for accurate prediction of field size factor in a varian probeam proton PBS system *J. Appl. Clin. Med. Phys.* **21** 197–204

Hussein M, Clark C H and Nisbet A 2017 Challenges in calculation of the gamma index in radiotherapy—towards good practice *Phys. Medica* **36** 1–11

ICRU report 90 2016 *Key Data for Ionizing-Radiation Dosimetry: Measurement Standards and Applications*

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

Magro G, Molinelli S, Mairani A, Mirandola A, Panizza D, Russo S, Ferrari A, Valvo F, Fossati P and Ciocca M 2015 Dosimetric accuracy of a treatment planning system for actively scanned proton beams and small target volumes: monte carlo and experimental validation *Phys. Med. Biol.* **60** 6865–80

Mohan R 2022 A review of proton therapy - current status and future directions *Precis. Radiat. Oncol.* **6** 164–76

Newhauser W, Fontenot J, Zheng Y, Polf J, Titt U, Koch N, Zhang X and Mohan R 2007 Monte Carlo simulations for configuring and testing an analytical proton dose-calculation algorithm *Phys. Med. Biol.* **52** 4569–84

Nocedal J and Wright S 2006 *Numerical optimization* (New York: Springer New York) 2nd edition978-0387-30303-1

Paganetti H 2012 Range uncertainties in proton therapy and the role of Monte Carlo simulations *Phys. Med. Biol.* **57** 99–117

Paganetti H, Jiang H, Parodi K, Slopsema R and Engelsman M 2008 Clinical implementation of full Monte Carlo dose calculation in proton beam therapy *Phys. Med. Biol.* **53** 4825–53

Particle Therapy Co-Operative Group (PTCOG) Online: https://ptcog.site/index.php/facilities-in-operation-public

Perl J, Shin J, Schümann J, Faddegon B and Paganetti H 2012 TOPAS: An innovative proton Monte Carlo platform for research and clinical applications *Med. Phys.* **39** 6818–37

Rahman M, Bruza P, Lin Y, Gladstone D J, Pogue B W and Zhang R 2020 Producing a Beam model of the varian probeam proton therapy system using topas monte carlo toolkit *Med. Phys.* **47** 6500–8

RaySearch Laboratories, Fast and accurate dose computations Online: https://raysearchlabs.com/fast-and-accurate-dose-computations/

RaySearch 2019 *RayStation 9B Reference Manual* (Stockholm, Sweden: RaySearch Laboratories AB)

Saini J, Maes D, Egan A, Bowen S R, St James S, Janson M, Wong T and Bloch C 2017 Dosimetric evaluation of a commercial proton spot scanning Monte-Carlo dose algorithm: Comparisons against measurements and simulations *Phys. Med. Biol.* **62** 7659–81

Saini J, Traneus E, Maes D, Regmi R, Bowen S R, Bloch C and Wong T 2018 Advanced Proton Beam Dosimetry Part I: Review and performance evaluation of dose calculation algorithms *Transl. Lung Cancer Res.* **7** 171–9

Sarrut D *et al* 2014 A review of the use and potential of the GATE Monte Carlo simulation code for radiation therapy and dosimetry applications *Med. Phys.* **41** 064301

Sawakuchi G O, Titt U, Mirkovic D, Ciangaru G, Zhu X R, Sahoo N, Gillin M T and Mohan R 2010 Monte Carlo investigation of the low-dose envelope from scanned proton pencil beams *Phys. Med. Biol.* **55** 711–21

Schreuder A N, Bridges D S, Rigsby L, Blakey M, Janson M, Hedrick S G and Wilkinson J B 2019 Validation of the RayStation Monte Carlo dose calculation algorithm using realistic animal tissue phantoms *J. Appl. Clin. Med. Phys.* **20** 160–71

Smith E A K, Winterhalter C, Underwood T S A, Aitkenhead A H, Richardson J C, Merchant M J, Kirkby N F, Kirby K J and Mackay R I 2022 A Monte Carlo study of different LET definitions and calculation parameters for proton beam therapy *Biomed. Phys. Eng. Express* **8** 015024

Tommasino F, Fellin F, Lorentini S and Farace P 2018 Impact of dose engine algorithm in pencil beam scanning proton therapy for breast cancer *Phys. Medica* **50** 7–12

Tourovsky A, Lomax A J, Schneider U and Pedroni E 2005 Monte Carlo dose calculations for spot scanned proton therapy *Phys. Med. Biol.* **50** 971–81

Trnková P, Bolsi A, Albertini F, Weber D C and Lomax A J 2016 Factors influencing the performance of patient specific quality assurance for pencil beam scanning IMPT fields *Med. Phys.* **43** 5998–6008

Varian Medical Systems 2020 *Varian Eclipse Proton Algorithms Reference Guide - ProBeam* (Palo Alto, CA, USA: Varian Medical Systems)

Varian Medical Systems: Treatment Planning for the ProBeam System Online: https://varian.com/en-au/products/proton-therapy/treatment-planning-probeam-system

Verburg J M, Grassberger C, Dowdell S, Schuemann J, Seco J and Paganetti H 2016 Automated Monte Carlo Simulation of Proton Therapy Treatment Plans *Technol. Cancer Res. Treat.* **15** NP35–46

Vlachoudis V 2009 Flair: A powerful but user friendly graphical interface for FLUKA *International Conference on Mathematics, Computational Methods & Reactor Physics (Saragota Springs, New York, 2009)* (American Nuclear Society) 790–800

Waters L S, McKinney G, Durkee J *et al* 2007 The MCNPX monte carlo radiation transport code *Hadronic Shower Simulation Workshop (Batavia, Illinois (USA), 2006)* (AIP Conference Proceedings) 81–90

Winterhalter C, Aitkenhead A, Oxley D, Richardson J, Weber D C, MacKay R I, Lomax A J and Safai S 2020a Pitfalls in the beam modelling process of Monte Carlo calculations for proton pencil beam scanning *Br. J. Radiol.* **93** 20190919

Winterhalter C *et al* 2018 Validating a monte carlo approach to absolute dose quality assurance for proton pencil beam scanning *Phys. Med. Biol.* **63** 175001

Winterhalter C *et al* 2020b Evaluation of GATE-RTion (GATE/Geant4) Monte Carlo simulation settings for proton pencil beam scanning quality assurance *Med. Phys.* **47** 5817–28

Xu Y, Zhang K, Liu Z, Liang B, Ma X, Ren W, Men K and Dai J 2022 Treatment plan prescreening for patient-specific quality assurance measurements using independent Monte Carlo dose calculations *Front. Oncol.* **12** 1–10

Yeom Y S *et al* 2020 A Monte Carlo model for organ dose reconstruction of patients in pencil beam scanning (PBS) proton therapy for epidemiologic studies of late effects *J. Radiol. Prot.* **40** 225–42

Yepes P, Adair A, Grosshans D, Mirkovic D, Poenisch F, Titt U, Wang Q and Mohan R 2018 Comparison of Monte Carlo and analytical dose computations for intensity modulated proton therapy *Phys. Med. Biol.* **63** 045003

Zhao L 2013 Effect of dose calculation grid size on proton dose calculation *Med. Phys. Fifty-fifth annual meeting of the American association of physicists in medicine* 40, 338