Zeitschrift für Medizinische Physik xxx (xxxx) xxx

Contents lists available at ScienceDirect

Zeitschrift für Medizinische Physik

journal homepage: www.elsevier.com/locate/zemedi

Original Paper

Error detection sensitivity and operational efficiency of phantom-based and phantom-less patient-specific quality assurance in proton therapy

Lukas Cornelius Wolter [a,b], Yazeed Ghannam [a], Kenneth Poels [c], Stefan Menkel [d], Fabian Hennings [a,b], Kevin Souris [e], Theresa Lenk [d], Kristin Stützer [a,b], Christian Richter [b,d,f,*]

[a] OncoRay – National Center for Radiation Research in Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Helmholtz-Zentrum Dresden - Rossendorf, Dresden, Germany
[b] Helmholtz-Zentrum Dresden - Rossendorf, Institute of Radiooncology - OncoRay, Dresden, Germany
[c] Department of Radiotherapy Oncology, University Hospitals Leuven, Leuven, Belgium
[d] Department of Radiotherapy and Radiation Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany
[e] Ion Beam Applications SA, Louvain-la-Neuve, Belgium
[f] German Cancer Consortium (DKTK), Partner Site Dresden, and German Cancer Research Center (DKFZ), Heidelberg, Germany

ARTICLE INFO

Keywords:
Sensitivity analysis
Patient-specific QA
Proton therapy
Automation

ABSTRACT

Background and purpose: In proton therapy (PT), patient-specific quality assurance (PSQA) is an important component of the measures which ensure accurate and safe treatment delivery. Traditional phantom-based PSQA is resource-intensive and may miss clinically relevant data transfer and delivery errors. This study compared the sensitivity and operational effort of established phantom-based methods versus an automated, phantom-less alternative based on the increasingly utilized log file-based approach.

Materials and methods: We evaluated phantom-based (specifically including dose measurements & manual physics checks) and phantom-less (specifically based on log file-based QA & automated physics checks) PSQA workflows. Twenty-nine artificial error scenarios were introduced to a clinically delivered head-and-neck plan. Error detection sensitivity was determined via the percentage of scenarios detected by each workflow. Operational effort was quantified by counting mouse clicks and manual parameter entries in clinical systems.

Results: Phantom-less PSQA detected 90% of simulated errors, outperforming phantom-based PSQA detecting only 52% at clinically realistic measurement conditions. Specifically, log file-based QA alone detected 83% of scenarios versus 10% detected by phantom-based measurements. Manual and automated plan parameter checks showed an equal sensitivity, detecting 48% of scenarios. The more automated, phantom-less alternative could reduce operational effort by at least one third of the currently required mouse clicks.

Conclusions: Phantom-less PSQA based on log file-based QA and automated physics checks provided higher error detection sensitivity and significantly reduced manual operational effort compared to conventional phantom-based and manual methods. These findings support its integration into clinical practice, a key objective of many PT centers.

1. Introduction

To leverage the full potential of pencil beam scanning (PBS) proton therapy (PT) [1,2] while ensuring patient safety, rigorous quality assurance (QA) is required. In addition to machine-QA performed daily to annually [3], every treatment plan undergoes patient-specific QA (PSQA) to verify three critical aspects [4]: (1) correctness of dose calculation by the treatment planning system (TPS), (2) integrity of plan data transferred between clinical systems, and (3) accurate delivery of the treatment plan by the PT system. Traditionally, these aspects are implicitly verified by re-calculating and measuring the treatment plan dose in water-equivalent phantoms [5]. Although widely established, phantom-based PSQA is resource-intensive, requiring additional beam time and expert personnel.

* Corresponding author at: Fetscherstr. 74, PF 41, 01307 Dresden, Germany
E-mail address: christian.richter@oncoray.de (C. Richter).

https://doi.org/10.1016/j.zemedi.2026.04.002
Received 7 November 2025; Accepted 17 April 2026

The transition from phantom-based, manual PSQA to phantom-less, automated solutions was identified as a major objective at the ESTRO Physics Workshop in 2023. Phantom-less workflows aim to explicitly address the three core aspects of PSQA. (1) TPS dose calculation can be verified using an independent dose calculation system [6]. While Monte Carlo is considered the gold-standard, faster analytical models may be suitable in selected cases [7]. (2) Data integrity can be ensured if the exported plan parameters match with the data before export, which can be accomplished by comparing TPS data with DICOM files of the treatment plan or machine steering files [8] at different workflow stages. (3) Delivery feedback from the PT system is accessible via log files of nozzle-internal monitor chambers, enabling detailed analysis of delivered spot parameters and also dose reconstruction on the patient-CT [9–11]. Specifically, log file-based QA (LFQA) acts as the primary driver of automated, phantom-less PSQA workflows, representing a reliable alternative to phantom-based measurements.

In photon therapy, phantom-less PSQA workflows have been widely adopted and are commercially supported for IMRT and VMAT. In contrast, its clinical translation in PT remains at an earlier stage. Although commercial solutions are now available, published data on real-world commissioning, sensitivity benchmarking, and operational impact of phantom-less PSQA in PT are still scarce, limiting evidence-based clinical adoption. Marmitt et al. [12] developed an in-house QA platform integrating the MCsquare dose engine [6] with LFQA, while Albertini et al. [13] clinically implemented the first daily-adaptive PT workflow, featuring independent dose calculation based on machine steering and delivery log files on the patient-CT, with automated structure and plan parameter checks. While these studies demonstrate technical feasibility, systematic benchmarking of error detection sensitivity against established phantom-based PSQA under clinically relevant error scenarios is lacking.

Such benchmarking is particularly timely, as international efforts are underway to formalize recommendations for PSQA in PT. The European Particle Therapy Network (EPTN) of ESTRO is currently developing a dedicated PSQA guideline which is expected to be finalized in 2026. In parallel, AAPM TG-201 explicitly recommends evaluating QA performance using clinically realistic error scenarios [14]. Quantitative sensitivity data, as provided in this study, are therefore helpful to support guideline development and verify that phantom-less PSQA offers comparable or superior performance compared to traditional phantom-based protocols.

Beyond sensitivity, workflow efficiency needs to be considered during clinical commissioning, especially in PT, where PSQA is subject to stringent regulatory oversight and requires additional personnel resources in form of highly specialized staff. Phantom-based PSQA typically involves manual phantom setup, beam delivery, data handling, and documentation, resulting in substantial human workload and machine occupancy. Reducing these efforts has direct implications for staff capacity, operational cost, and patient throughput, making operational effort analysis an essential endpoint.

This study aimed to quantify the effectiveness and efficiency of phantom-based and phantom-less PSQA workflows. First, we assessed the error detection capabilities of two clinical phantom-based PSQA approaches and one phantom-less PSQA prototype based on artificial error scenarios (sensitivity analysis). Second, the status-quo workload to complete phantom-based PSQA procedures was measured in mouse clicks required for all human operations in clinical systems, followed by an estimation of mouse click savings introduced by the automated, phantom-less alternative (operational effort analysis). To our knowledge, this is the first side-by-side assessment of error detection sensitivity and human workload between phantom-based and phantom-less PSQA approaches in PT.

Our findings contribute to the ongoing advancement of reproducible, vendor-independent QA solutions, which are not only important for pre-treatment PSQA but in particular for online-adaptive PT (OAPT).

2. Materials and methods

2.1. Treatment delivery system

The University Proton Therapy Dresden (UPTD, DE) houses an IBA Proteus Plus PT system (Ion Beam Applications, Louvain-la-Neuve, BE) with a universal nozzle. Pencil beams (100.0–226.7 MeV) are scanned by two scanning magnet pairs (x – fast-, y – slow-scanning direction in IEC-61217 beam's eye view coordinates). One monitor unit (MU) corresponded to an absorbed physical dose of 0.108–0.112 cGy, measured at a fixed depth in the proximal region of the depth-dose curve of a mono-energetic, quadratic PBS reference field, depending on energy. Treatment plans were created with RayStation v2023B (RaySearch Laboratories, Stockholm, SE) and prepared for delivery in the oncology information system (OIS) MOSAIQ v2.83 (Elekta AB, Stockholm, SE).

2.2. Investigated PSQA workflows

The PSQA process spans from clinical plan approval to delivery of the first treatment fraction. Currently, the pre-treatment phantom-based PSQA protocol at UPTD (Table 1) includes dose measurements in water-equivalent materials and manual plan parameter checks. The original plan is recalculated in water at 0° gantry and table angle for perpendicular field-by-field measurements [5] using two setups: (a) absolute point dose, measured with a Semiflex ionization chamber (IC; PTW, Freiburg, DE) in an IBA Blue water phantom and compared to TPS values with a ±4% tolerance; (b) relative 2D dose distributions, acquired with an IBA Lynx PT detector at multiple depths in solid water-equivalent material (RW3) and evaluated against the TPS using global 2D gamma index analysis (3%/3mm, 10% low dose cut-off) in OmniPro I'mRT (IBA), accepting gamma pass rates (GPRs) ≥95%. Measurements in (a) and (b) are conducted at predefined, plan-specific points/planes to capture the high-dose region of each PBS field. Phantom-based QA-plan measurements are prepared in the OIS after plan import from the TPS, including the creation and review of additional QA documents. Two independent certified medical physics experts (MPEs) perform manual plan parameter comparisons between TPS and OIS. This "four-eye" redundancy principle of the 1st and 2nd MPE check is a safety precaution set mandatory at UPTD.

An alternative, phantom-less PSQA workflow (Table 1) using LFQA and automated physics checks (APC) was implemented. In this prototype, all PBS fields were delivered to a beam stopper at planned gantry positions, without phantom-based measurements. A log file plan was generated from treatment delivery log files recorded by nozzle-internal ICs [15] using a Python script. For reconstruction of the delivered dose, the clinical TPS algorithm (RaySearch Proton Monte Carlo v5.5) was used, as the expected differences to an independent Monte Carlo dose calculation engine are not relevant for the presented analysis. Reconstructed doses were compared to the planned dose by means of global 3D gamma index analysis in VeriSoft (PTW) and point dose differences. Due to the absence of setup uncertainties and the high precision of recorded beam parameters, a stricter gamma criterion was selected (2%/2mm, 10% low dose cut-off, ≥95% GPR threshold). LFQA also included spot-by-spot comparison (difference histograms, spot map plots) to detect position/MU errors, allowing spot-wise deviations from the original plan up to 2 mm/0.05 MU, respectively. These tolerance criteria were based on a prior log file validation study that investigated the reproducibility of discrepancies between recorded and planned parameters, thereby supporting LFQA as a reliable and safe alternative to phantom-based measurements [15].

Data transfer was validated through script-based, zero-tolerance APC between the TPS- and OIS-exported DICOM RTPLANs, replicating the currently employed, manual MPE checks. These checks included machine parameters (e.g., gantry and table angle, air gap, range shifter) as well as beam parameters (e.g., spot count, monitor units, proton range). Both LFQA and APC featured automated QA report generation (cf.

Table 1
Comparison of phantom-based and phantom-less PSQA workflows at UPTD. Phantom-based measurements are replaced by LFQA and manual MPE checks have been automated.

	Phantom-based PSQA	Tolerance	Phantom-less PSQA	Tolerance
Delivery verification	Dose measurements: • Point dose (Semiflex) • 2D gamma index analysis on plane dose (Lynx)	 4% Γ(3%/3mm) > 95%	Log file-based QA: • Spot statistics (position, MU) • 3D gamma index analysis on log file-reconstructed dose	 2 mm, 0.05 MU Γ(2%/2mm) > 95%
Data transfer validation	Manual physics checks: • 1st MPE check • 2nd MPE check	 Exact match	Automated physics checks: • Script-based plan comparison based on MPE checks	 Exact match

Supplement 1 & 2, respectively).

2.3. Sensitivity analysis

A systematic assessment of both PSQA workflows' error detection sensitivity was conducted for one field of a clinically delivered head-and-neck cancer (HNC) plan (Fig. 1). The clinical target volume (CTV) was treated with a prescribed mean dose of 70 Gy (RBE) over 33 fractions. Three PBS fields were planned with gantry (G) and treatment table (T) angles 1: G180°/T0°, 2: G60°/T345°, 3: G300°/T15°, respectively.

To simulate planning errors or corrupted data transfer between clinical subsystems leading to flawed beam delivery by the PT system, eight types of plan parameter errors were introduced into field 2. The selected plan parameters have been identified as high-risk factors in internal and published failure mode and effect analyses [16]. Each error type was implemented at three to four severity levels, yielding a total of 29 error scenarios of the clinical reference plan. Severity levels I–IV were determined based on the relative dose deviation from the reference plan obtained by intentional plan parameter manipulation, with severity levels ≥III (local dose deviation ≥3%) being considered as clinically relevant (Table 2).

All scenario plans were subjected to the complete phantom-based and phantom-less PSQA workflow under ideal and clinically realistic conditions: Under ideal conditions, phantom-based measurements were taken at the TPS-identified location of maximum dose error for each scenario. Under realistic conditions, measurements used the clinical routine detector positions of the original QA plan. TPS-predicted scenario doses hereby served as a benchmark for dose differences and GPRs anticipated from phantom-based measurements and LFQA.

A scenario plan was considered detected if it violated any of the UPTD-specific tolerance criteria in measurements, LFQA, or plan parameter checks against the clinical plan. The error detection sensitivity of each PSQA method was defined as the relative number of detected versus total scenarios:

$$S[\%] = \frac{\#\ \text{detected scenarios}}{\#\ \text{total scenarios}} \times 100 \quad (1)$$

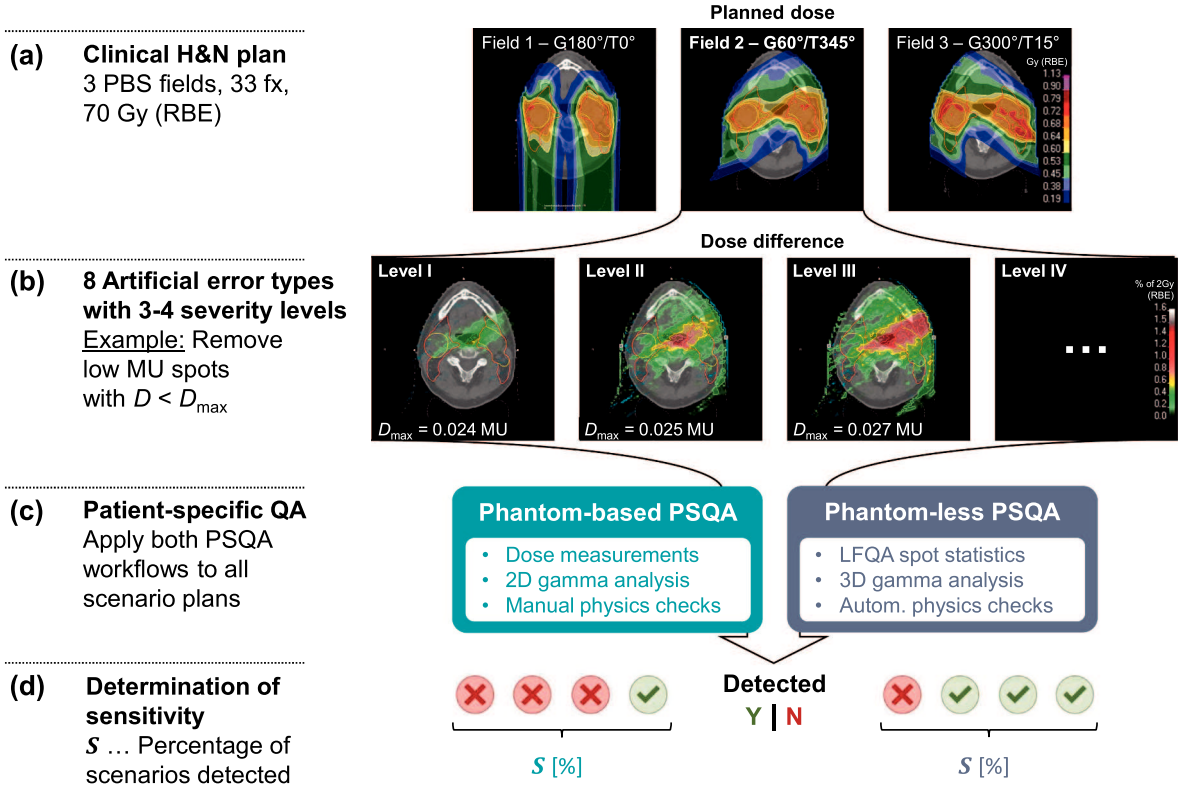

Fig. 1. Methodology of the PSQA error detection sensitivity analysis. Planned field dose distributions of the clinical HNC plan (a). Twenty-nine error scenarios were derived from field 2 (G60°/T345°) by manipulating selected beam parameters, causing dose differences of different severity to the reference plan (b). All scenario plans were subjected to both PSQA workflows (c). An error scenario was successfully detected if any PSQA passing criterion was violated. Sensitivity was defined as the relative number of detected error scenarios (d).

Table 2
Artificial error scenarios introduced to field 2 of the clinical reference plan by manipulating eight different beam parameters with I–IV severity levels each, based on relative local dose deviations from the clinical reference plan. Local dose deviations ≥3% (severity level III) were considered as clinically relevant. Severity level IV errors represent extreme dose deviations and were only implemented for selected scenarios.

No.	Scenario error type	Severity level (based on approximate dose deviation from reference plan)			
		I (<1%)	**II** (<3%)	**III** (≥3%)	**IV** (>5%)
1	Remove one spot with	0.14 MU	0.35 MU	0.79 MU	1.29 MU
2	Shift one spot in $+x$ direction by	1 mm	2 mm	4 mm	6 mm
3	Shift one row in $+x$ direction by	1 mm	2 mm	4 mm	6 mm
4	Shift one layer in $+x$ direction by	1 mm	2 mm	4 mm	–
5	Increase all MU by	1%	2%	4%	6%
6	Remove low MU spots of less than	0.024 MU	0.025 MU	0.027 MU	–
7	Increase air gap by	10 mm	20 mm	30 mm	–
8	Increase energy of one layer by	0.2 MeV	0.3 MeV	0.5 MeV	0.7 MeV

2.4. Operational effort analysis

The total duration of a full PSQA varies with plan complexity and operator experience, typically requiring a minimum of three hours. Limited beam time availability allows setup of only one phantom per QA session, extending the total workflow to at least two days in most cases. Phantom setup, irradiation and dismantling requires approximately 45 (Lynx) to 60 minutes (water). Both phantoms are assembled at least twice a week at UPTD.

We recorded the number of mouse clicks and manual entries of critical treatment parameters in clinical systems (TPS, OIS, external documentation) as a surrogate for the human effort required to complete PSQA for a single PBS field. This approach eliminated the influence of inter-operator and plan variability. The PSQA process was divided into four subprocesses: QA plan generation, OIS plan import, 1st and 2nd MPE check and phantom-based measurements. Phantom-based measurements have been excluded from this assessment, as their duration depends on setup and delivery rather than clicks. Double-clicks were treated as single clicks to avoid over-interpretation. We further compared the tasks of the phantom-based workflow with those in the phantom-less alternative to estimate its automation potential, as certain operations might be streamlined or even eliminated.

To contextualize the operational effort, we compared the pre-treatment PSQA process at UPTD with that of another clinical PT center, PARTICLE (UZ Leuven, BE) using mouse click analysis. The key difference between both centers was the OIS in use: UPTD used MOSAIQ, while PARTICLE used RayCare, which has a different level of integration with the RayStation TPS used in both centers. PARTICLE further utilized IBA's myQA iON to evaluate phantom-based measurements. This software also featured independent Monte Carlo dose calculation and LFQA, whose mouse clicks were excluded from the status-quo comparison to ensure fair workflow comparison. However, the clicks performed in myQA iON were used as a realistic workload estimation for LFQA in the phantom-less alternative.

For this investigation, the UPTD and PARTICLE workflows were referred to as "separate-OIS" and "integrated-OIS" workflow, respectively. The separate-OIS workflow was applied to a three-field HNC plan, with the mouse click count normalized to the count required for a single field. The integrated-OIS workflow was applied to a single-field esophagus plan.

For better clinical interpretability of the operational effort analysis, approximate time requirements for all four PSQA sub-processes were derived from repeated workflow executions by multiple MPEs over time. Both the sensitivity and operational effort analyses were designed to be independent of the anatomical treatment site, as PSQA workflow steps and error detection mechanisms are not affected by the target region directly and depend primarily on treatment plan complexity.

3. Results

3.1. Sensitivity analysis

Phantom-based PSQA detected 15 out of 29 error scenarios, yielding a combined sensitivity of 52% with measurements and manual MPE checks under clinically realistic PSQA conditions, where the same predefined measurement location was used for all scenario plans (Fig. 2). In contrast, the phantom-less approach detected 26 out of 29 errors, achieving a combined sensitivity of 90% with LFQA and APC. While both manual and automated physics checks exhibited an equal stand-alone sensitivity of 48%, LFQA outperformed phantom-based measurements, showing a stand-alone sensitivity of 83% compared to 10%, respectively.

While all error scenarios with severity level >1 were detectable with phantom-less PSQA, phantom-based PSQA failed to detect certain error types entirely. Specifically, only 7/13 scenarios of severity level ≥III, that would be considered clinically relevant due to local dose deviations ≥3%, were detected with phantom-based PSQA. Only when executed under ideal conditions, i.e., when performing measurements at the TPS-predicted locations of maximum dose deviation for each error scenario, phantom-based PSQA achieved the same combined sensitivity of 90%, thus detecting all clinically relevant errors.

Log file-reconstructed point doses agreed with Semiflex measurements within ±3.5% for each scenario, showing clear trends towards higher dose deviations from the reference plan with increasing severity level. Point dose measurements detected 17/29 scenarios at ideal conditions, yielding a stand-alone sensitivity of point dose measurements of 59%. At realistic conditions, only the highest global overdosage (scenario 5, severity level IV) could be detected with point dose measurements, resulting in a stand-alone sensitivity of only 3% (Supplement 3). For reference, point dose measurements of the unaltered clinical plan agreed with the planned dose within ±1%, reflecting realistic PSQA measurement conditions at a homogeneous location within the high-dose region.

Under ideal/realistic conditions, Lynx measurements detected three/two scenarios using the 2D Γ(3%/3mm) criterion for relative plane dose distributions, resulting in a stand-alone sensitivity of 10%/7%, respectively.

LFQA, on the other hand, evaluated the full 3D plan dose without depending on external detector placement, and hence did not distinguish between ideal and realistic conditions. Using the more sensitive 3D Γ(2%/2mm) criterion, it detected 8/29 scenarios (Supplement 3). Almost every scenario detected by 3D gamma index analysis (stand-alone sensitivity of 28%) was also detected through spot-by-spot comparison (stand-alone sensitivity of 76%), which means that the enhanced sensitivity of LFQA mostly stems from spot statistics rather than dose reconstruction for the investigated scenarios.

3.2. Operational effort analysis

Completing all four sub-processes of the separate-OIS workflow for one PBS field of the HNC plan required 315 mouse clicks (Fig. 3), with the majority used for importing the treatment and QA plans into the OIS. In comparison, the integrated-OIS workflow for the esophagus plan required 255 clicks, primarily for performing manual MPE checks.

The separate-OIS workflow further demanded 18 manual entries of critical treatment parameters (e.g., patient name, prescription, range shifter, MU), while 10 such operations were recorded for the integrated-OIS workflow. Clicks required for 1st and 2nd MPE check were

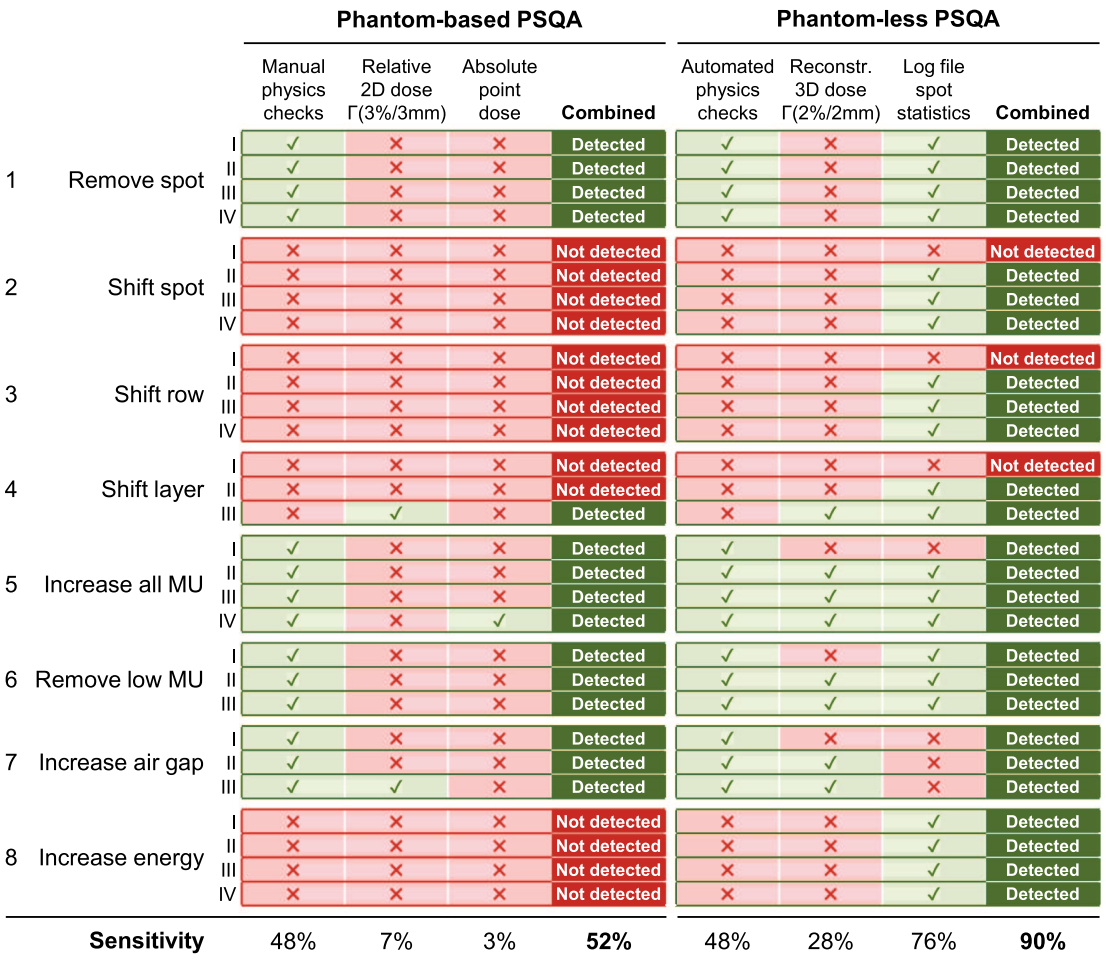

Fig. 2. Detailed breakdown of artificial error scenarios detected by both PSQA methods under clinically realistic conditions. In phantom-based PSQA (left), most sensitivity arises from the 1st and 2nd MPE checks rather than dose measurements. Phantom-less PSQA (right) benefits from log file information depth enabling 3D dose evaluation and spot-level statistical analysis. The implementation of the stricter 2%/2mm gamma criterion for log file-reconstructed dose distributions resulted in only minor sensitivity gains.

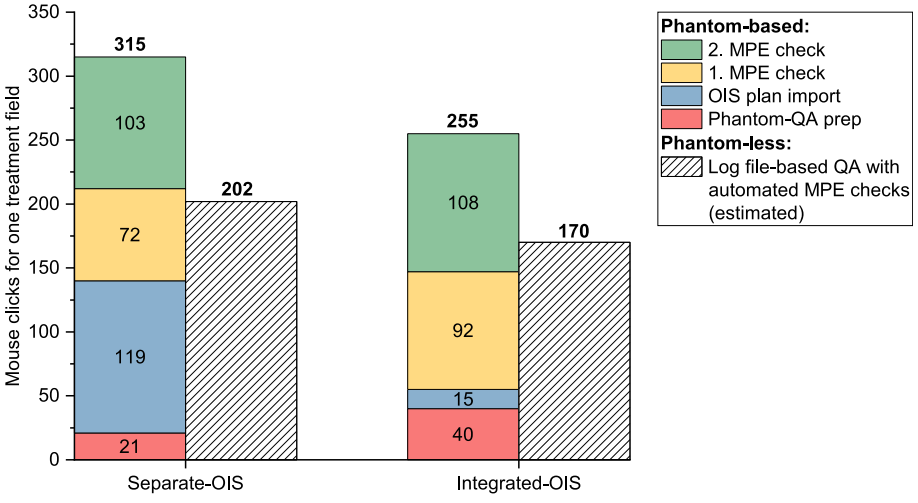

Fig. 3. Mouse clicks required to complete all investigated sub-processes of the phantom-based separate-OIS (RayStation + MOSAIQ) and integrated-OIS (RayStation + RayCare) PSQA workflow for a single-field PBS treatment plan (colored bars). Potential click reduction was estimated by evaluating the necessity of mouse clicks currently required in the phantom-based workflow for the alternative, phantom-less workflow utilizing LFQA and APC (hatched bars).

comparable across both workflows, but large differences arose for QA plan generation and OIS plan import. In the integrated-OIS workflow, QA-plan generation effort was almost doubled, while OIS plan import required 87% less clicks.

For the current separate-OIS workflow, QA plan preparation required approximately 60 min per treatment plan. Importing the QA

plan into the OIS required an additional 30 min. The 1st and 2nd MPE check both required 30–45 min, depending on plan complexity and operator experience. Taken together, all four sub-processes accounted for approximately 150–180 min per treatment plan.

For the integrated-OIS workflow, corresponding execution times for QA plan preparation, OIS plan import, and first and second MPE checks were recorded analogously. The approximate durations were: QA plan preparation: 10 min, OIS plan handling: 10 min, first and second MPE check: 30–60 min, resulting in a total duration of approximately 50 min–80 min per treatment plan.

The implementation of a phantom-less alternative workflow could potentially reduce the currently required mouse clicks to 202 (–36%) in the separate-OIS workflow and 170 (–33%) in the integrated-OIS workflow, either through task automation (e.g., utilizing APC instead of redundant MPE checks) or deprecation (e.g., eliminating QA plan generation with LFQA). The total duration of phantom-less PSQA amounts to approximately 50 min per treatment plan, comprising 45 min for LFQA including independent dose calculation in myQA iON, and 5 min for APC execution. This translates to a PSQA time reduction of up to 72% for the separate-OIS workflow and 38% for the integrated-OIS workflow, without considering time savings due to the absence of phantom setup, measurements and analysis.

4. Discussion

This study quantified the advantages of phantom-less over phantom-based PSQA in PT, with an automated, log file-based strategy demonstrating superior error detection sensitivity and workflow efficiency. By exposing the inherent limitations of current phantom-based solutions, our findings point toward the necessity to redesign QA processes and adopt an automated, phantom-less approach. The simultaneous assessment of sensitivity and operational burden represents a novel and practical contribution to this ongoing paradigm shift: Beyond validating a specific solution, we provided a reproducible strategy for commercial PT system environments to benchmark and commission alternative phantom-less QA workflows, facilitating broader clinical adoption through objective performance metrics.

Applying both PSQA methods to artificial error scenarios, we found that the phantom-less approach effectively detected even subtle single-spot deviations thanks to log file information enabling spot-level statistics and 3D gamma index analysis. In contrast, phantom-based QA, limited to point dose measurements and 2D gamma index analysis, struggled to detect such errors, especially under clinically realistic conditions. These differences arose from inherent limitations of phantom-based setups, where dose was measured at predefined, plan-specific points/planes with high tolerances. As a result, the likelihood of capturing an out-of-tolerance deviation was low, and some clinically relevant dose errors remained undetected.

In particular, the phantom-based PSQA workflow failed to detect the majority of scenarios involving spot position and energy perturbations. Depending on the number of affected spots and the magnitude of the deviation, these error types can result in substantial dose distortions and were associated with multiple undetected scenarios classified as clinically relevant (severity levels III and IV). Conversely, the phantom-less approach only failed to detect very small spot position deviations limited to severity level I. These scenarios were intentionally introduced as minimal perturbations to challenge both PSQA workflows and were not expected to result in clinically relevant dose effects, even when applied to multiple spots.

Inconsistencies between measured doses and their TPS predictions often resulted from detector positioning uncertainties at steep dose gradients, especially when point dose alterations were introduced. As ideal measurement locations coincided with those of the most pronounced dose gradients, an agreement of planned and Semiflex-measured point doses within $\pm 3.5\%$ is considered satisfactory. While detector positioning affected both Semiflex and Lynx measurements, 2D GPRs from Lynx measurements may have been further influenced by dose normalization effects.

The more global 3D gamma index analysis in the phantom-less workflow overcomes many of these issues by eliminating the reliance on predefined measurement locations. LFQA provides more clinically relevant feedback by reconstructing dose on the patient-CT under treatment conditions (planned gantry/couch angle) – in contrast to the simplified QA geometries in phantom-based workflows (angles collapsed to 0°). Sensitivity could be further improved by separately evaluating high- and low-dose regions or applying local gamma criteria. The key advantage of LFQA, however, was its ability to perform spot-level statistical analysis, which effectively identified nearly all single-spot manipulations. Nevertheless, inherent beam measurement uncertainties of the nozzle-internal ICs require non-zero tolerances in LFQA, meaning that small variations in spot position or monitor units within these thresholds may go undetected. These tolerance levels, however, were set sufficiently low to ensure that any undetected deviations could be confidently regarded as clinically irrelevant. In contrast, changes in air gap – which are typically of clinical relevance – cannot be obtained from log file data, as the snout position is not recorded. Such discrepancies, however, were captured by the physics checks, demonstrating the complementary synergy between LFQA and APC.

Since log files can be susceptible to PT system-intrinsic errors, the reproducibility of log file-derived spot parameters and dose has therefore been thoroughly evaluated across multiple fractions of clinical treatment plans beforehand, demonstrating negligible deviations from planned values [15]. In addition, delivery verification of spot position, size and MU is already part of the standard daily QA at most PT centers. These existing procedures could easily be extended to include automated log file analyses of daily QA deliveries, thereby providing a continuous validation mechanism (QA of log file data) that safeguards the full reliance on log files instead of phantom-based measurements for PSQA.

Of note, phantom-based PSQA is usually replaced with a combination of LFQA and an independent secondary dose calculation, either within a log file-based dose reconstruction or based on the original treatment plan. This combination is also favored in current discussions within the European Particle Therapy Network (EPTN) to develop an ESTRO guideline for PSQA in particle therapy.

Both plan parameter check solutions showed identical detection performance, as the automated method was a script-based replication of manual MPE checks. While this did not enhance sensitivity, it improved efficiency by eliminating the need for redundant four-eye evaluation by two MPEs. At the time of this study, MPE checks were limited in scope, with only the first and last layer energies being reviewed, and individual spots not being assessed. Consequently, changes to intermediate layer energies or single spot parameters remained unnoticed.

Our findings were consistent with previous reports by Matter et al. [17], who demonstrated superior sensitivity of phantom-less PSQA using log file dose reconstruction and zero-tolerance checks on machine steering files, achieving 100% error detection in their study. Similar sensitivity levels could have been achieved by adding explicit spot-by-spot deviation checks to our DICOM-based APC, without increasing QA time. The sensitivity metrics used in both studies are not generalizable, as detection rates inherently depend on the specific types and severity levels of the simulated errors. Different error scenarios introduced to a different clinical reference plan may have yielded other detection performance for the same QA method. Nonetheless, our method remains valid for benchmarking other PSQA workflows under comparable conditions and is extensible to other PT centers and treatment sites.

From an operational effort perspective, the integrated-OIS phantom-based workflow involved fewer interactions due to better TPS-OIS integration and reduced time spent on manual report handling. More clicks were required for QA plan generation; however, this step took

more time to complete in the separate-OIS workflow. The increased time was mainly attributable to non-click interactions, including manual identification of homogeneous high-dose regions for Semiflex measurements, manual export of plane doses for Lynx measurements, and manual data entry into Excel spreadsheets.

Implementing the phantom-less alternative workflow reduced manual tasks by eliminating QA-plan generation, phantom-based measurement preparation and analysis, redundant MPE checks, as well as duplicate documentation steps. This could ease staff workload and decrease the likelihood of human error [16]. In addition, LFQA enables immediate error source identification at the single-spot level, offering a key efficiency gain over external dose measurements.

While mouse clicks offered an objective measure of PSQA operational effort in our study, they did not account for non-click tasks and cognitive load such as scrolling through images, interpreting data visually, typing and waiting. The observed reduction of 33–36% in mouse clicks is therefore a conservative estimate, considering the approximate time reduction of 38–72% achievable per treatment plan. Although not captured in the primary click-based operational effort analysis, phantom setup and irradiation times would also be eliminated in the phantom-less workflow, freeing up at least 2 h of treatment room occupancy per week. Similar efficiency gains have been reported by Hernandez-Morales et al. [18], who demonstrated that workflow duration reductions of at least 50% may be achieved just by partial automation of phantom-based operations. Combining such automation with LFQA, as presented in our study, may further reduce the workload by completely removing the need for phantom-based beam measurements.

In summary, this work highlighted the clinical potential of phantom-less PSQA as a more sensitive, efficient, and scalable alternative to traditional phantom-based methods in proton therapy. Our analysis demonstrated that phantom-less alternatives offer superior error detection sensitivity and reduced operational burden. For clinical translation, phantom-less PSQA could be implemented alongside existing phantom-based workflows during a validation phase, requiring virtually only minor extra operational effort. This parallel approach would help clinics to build confidence and gather site-specific evidence, facilitating a safe transition to fully phantom-less PSQA.

CRediT authorship contribution statement

Lukas Cornelius Wolter: Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Methodology, Formal analysis, Data curation. **Yazeed Ghannam:** Investigation, Formal analysis, Data curation. **Kenneth Poels:** Writing – review & editing, Resources, Investigation. **Stefan Menkel:** Writing – review & editing, Supervision, Methodology. **Fabian Hennings:** Writing – review & editing, Visualization. **Kevin Souris:** Writing – review & editing, Conceptualization. **Theresa Lenk:** Investigation. **Kristin Stützer:** Writing – review & editing, Visualization. **Christian Richter:** Writing – review & editing, Supervision, Project administration, Methodology, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: OncoRay has an institutional research agreement with, and Kevin Souris is an employee of, Ion Beam Applications S/A. The authors have no conflicts of interest to disclose.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.zemedi.2026.04.002.

References

[1] Jakobi A, Bandurska-Luque A, Stützer K, Haase R, Löck S, Wack L-J, et al. Identification of patient benefit from proton therapy for advanced head and neck cancer patients based on individual and subgroup normal tissue complication probability analysis. Int J Radiation Oncology*Biology*Physics 2015;92:1165–74. https://doi.org/10.1016/j.ijrobp.2015.04.031.

[2] Kooy HM, Grassberger C. Intensity modulated proton therapy. Br J Radiol 2015;88. https://doi.org/10.1259/bjr.20150195.

[3] Arjomandy B, Taylor P, Ainsley C, Safai S, Sahoo N, Pankuch M, et al. AAPM task group 224: Comprehensive proton therapy machine quality assurance. Med Phys 2019;46:e678–705. https://doi.org/10.1002/mp.13622.

[4] Zhu XR, Li Y, Mackin D, Li H, Poenisch F, Lee AK, et al. Towards effective and efficient patient-specific quality assurance for spot scanning proton therapy. Cancers (Basel) 2015;7:631–47. https://doi.org/10.3390/cancers7020631.

[5] Miften M, Olch A, Mihailidis D, Moran J, Pawlicki T, Molineu A, et al. Tolerance limits and methodologies for IMRT measurement-based verification QA: Recommendations of AAPM Task Group No. 218. Med Phys 2018;45. https://doi.org/10.1002/mp.12810.

[6] Souris K, Lee JA, Sterpin E. Fast multipurpose Monte Carlo simulation for proton therapy using multi- and many-core CPU architectures. Med Phys 2016;43:1700–12. https://doi.org/10.1118/1.4943377.

[7] Nenoff L, Matter M, Jarhall AG, Winterhalter C, Gorgisyan J, Josipovic M, et al. Daily Adaptive Proton Therapy: Is it Appropriate to Use Analytical Dose Calculations for Plan Adaption? Int J Radiat Oncol Biol Phys 2020;107:747–55. https://doi.org/10.1016/j.ijrobp.2020.03.036.

[8] Albertini F, Matter M, Nenoff L, Zhang Y, Lomax A. Online daily adaptive proton therapy. Br J Radiol 2020;93. https://doi.org/10.1259/bjr.20190594.

[9] Belosi MF, van der Meer R, Garcia de Acilu Laa P, Bolsi A, Weber DC, Lomax AJ. Treatment log files as a tool to identify treatment plan sensitivity to inaccuracies in scanned proton beam delivery. Radiotherapy and Oncology 2017;125:514–9. https://doi.org/10.1016/j.radonc.2017.09.037.

[10] Li H, Sahoo N, Poenisch F, Suzuki K, Li Y, Li X, et al. Use of treatment log files in spot scanning proton therapy as part of patient-specific quality assurance. Med Phys 2013;40:021703. https://doi.org/10.1118/1.4773312.

[11] Matter M, Nenoff L, Marc L, Weber DC, Lomax AJ, Albertini F. Update on yesterday's dose—Use of delivery log-files for daily adaptive proton therapy (DAPT). Phys Med Biol 2020;65:195011. https://doi.org/10.1088/1361-6560/ab9f5e.

[12] Guterres Marmitt G, Pin A, Ng Wei Siang K, Janssens G, Souris K, Cohilis M, et al. Platform for automatic patient quality assurance via Monte Carlo simulations in proton therapy. Phys Med 2020;70:49–57. https://doi.org/10.1016/j.ejmp.2019.12.018.

[13] Albertini F, Czerska K, Vazquez M, Andaca I, Bachtiary B, Besson R, et al. First clinical implementation of a highly efficient daily online adapted proton therapy (DAPT) workflow. Phys Med Biol 2024;69:215030. https://doi.org/10.1088/1361-6560/ad7cbd.

[14] Siochi RA, Balter P, Bloch CD, Santanam L, Blodgett K, Curran BH, et al. Report of Task Group 201 of the American Association of Physicists in Medicine: Quality management of external beam therapy data transfer. Med Phys 2021;48. https://doi.org/10.1002/mp.14868.

[15] Wolter LC, Hennings F, Bokor J, Richter C, Stützer K. Validity of one-time phantomless patient-specific quality assurance in proton therapy with regard to the reproducibility of beam delivery. Med Phys 2025. https://doi.org/10.1002/mp.17637.

[16] Cantone MC, Ciocca M, Dionisi F, Fossati P, Lorentini S, Krengli M, et al. Application of failure mode and effects analysis to treatment planning in scanned proton beam radiotherapy. Radiat Oncol 2013;8:1–9. https://doi.org/10.1186/1748-717X-8-127.

[17] Matter M, Nenoff L, Meier G, Weber DC, Lomax AJ, Albertini F. Alternatives to patient specific verification measurements in proton therapy: a comparative experimental study with intentional errors. Phys Med Biol 2018;63. https://doi.org/10.1088/1361-6560/aae2f4.

[18] Hernandez Morales D, Shan J, Liu W, Augustine KE, Bues M, Davis MJ, et al. Automation of routine elements for spot-scanning proton patient-specific quality assurance. Med Phys 2019;46:5–14. https://doi.org/10.1002/mp.13246.