

RESEARCH ARTICLE

Independent secondary dose calculation for patient-specific quality assurance: Quantitative benefit of Monte–Carlo and custom beam modeling

Lone Hoffmann^{1,2}  | Mai-Britt Linaa¹ | Ditte Sloth Møller^{1,2}

¹Department of Oncology, Aarhus University Hospital, Aarhus, Denmark

²Department of Clinical Medicine, Faculty of Health, Aarhus University, Aarhus, Denmark

Correspondence

Lone Hoffmann, Department of Oncology, Aarhus University Hospital, Aarhus, Denmark.
Email: Lone.Hoffmann@aarhus.rm.dk

Abstract

Background: Independent secondary dose calculation (ISDC) is becoming increasingly important for patient specific quality assurance. The most widely used analytical algorithms in ISDC are becoming challenged by Monte Carlo systems, which offer a potentially higher accuracy.

Purpose: Quantify the benefit of Monte Carlo over analytical algorithms, and of customized beam models over generic beam models, in terms of clinically relevant parameters, action level, and workload.

Methods: A set of 100 patients across 20 case classes, all planned with Acuros XB (Siemens Healthineers) was analyzed with Mobius3D (M3D) (Siemens Healthineers) and SciMoCa (Radiologica LLC), both with custom beam models (SMCcbm) and generic beam models (SMCgen). Gamma pass rate (GPR) and mean target dose difference $|\Delta D|$ action levels were determined for various rates of QA failures.

Results: At a workload of $< 10\%$, the action level for M3D was GPR (3%, 3 mm) $< 90\%$ and $|\Delta D| > 4.5\%$. For SMCgen, the action level was GPR (2%, 2 mm) $< 95\%$ and $|\Delta D| > 1.5\%$. For SMCcbm, it was GPR (2%, 1 mm) $< 95\%$ and $|\Delta D| > 1\%$. The combination of both criteria reduced the workload to $< 5\%$. SMC failures could be traced back to differences in the patient density model of Acuros XB. Some M3D failures could be traced back to the handling of tissue heterogeneities. The different performance between SMCcbm and SMCgen was due to one (of three) generic beam models.

Conclusion: Monte Carlo allows substantially stricter acceptance criteria and is sensitive enough to capture TPS commissioning errors. Generic beam models must be validated thoroughly before being put to use in ISDC.

KEYWORDS

action level, dose computation, independent secondary dose computation, Monte Carlo, patient specific quality assurance

1 | Introduction

The constant evolution and diversification of radiotherapy treatments has led to an ever-growing complexity of workflows and treatment planning tools, requiring

effective patient-specific quality assurance (PSQA). One such QA method of the treatment planning process, independent secondary dose calculation (ISDC), which incorporating accurate dose computation algorithms, has been recommended by the ICRU 83 and the AAPM

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 The Author(s). *Journal of Applied Clinical Medical Physics* published by Wiley Periodicals LLC on behalf of American Association of Physicists in Medicine.

TG 219 reports.^{1,2} The premise of effective ISDC is two-fold: (I) the accuracy of the employed algorithms needs to match or exceed that of the treatment planning software;¹ (II) the treatment plan needs to be validated via clinically relevant criteria defined for the patient anatomy.

At first glance, a more accurate and meticulously commissioned ISDC software raises the sensitivity to detect errors, such as overly modulated treatment plans, problematic TPS commissioning or faults in the construction of the patient density model. The sensitivity of a QA method is also a consequence of the action level, which defines what can be seen as a tolerable slip and what constitutes a significant incident. In practice, a picky action level results in a gratuitously high workload by producing false positives, whereas a more lenient one could risk serious mistreatment. More often than not, the rate of false positives—and thereby the workload—dictates—a workable action level.

False positives are caused by the inaccuracy of the software solutions employed in both the treatment planning system (TPS) and ISDC. If the planned and recalculated dose distribution disagree, it could be a foible of the TPS, one of the ISDC, or both. Hence, the most effective means to cut the rate of false positives is high overall accuracy of the ISDC, supposing that the TPS operates at its optimum. If it does not, a highly accurate and well commissioned ISDC may reveal the TPS as the cause of frequent false positives at a sensitive action level.

The question we explore here is how the inaccuracy of an ISDC bears on a workable action level. In theory, the action level can be derived from measurements of sensitivity and specificity, which requires the knowledge of the true fails in the samples.³ In radiotherapy practice, such ground truth is usually inaccessible (or established by decree). Hence, we suggest the following method: for a representative collection of clinical cases, first establish the action level that results in a certain plan failure rate. Then, compare various ISDC methods via the action level they permit for a given failure rate. In other words, we search for the most sensitive action level, that is, still compatible with clinical circumstances.

The ISDC solutions compared here are deliberately chosen to represent different concepts. The Mobius3D (Varian Medical Systems, Palo Alto, USA) software incorporates a convolution–superposition dose computation algorithm with generic beam models for the Varian TrueBeam linear accelerators employed in this study (M3D).⁴ The SciMoCa (Radialogica LLC, St. Louis, USA) software comprises a Monte Carlo algorithm paired with either generic (SMCgen) or custom-commissioned beam models (SMCcbm).^{5,6} Both products support gamma evaluation and various dose-volume metrics for plan analysis. This allows us to quantify the difference between a type B algorithm (M3D) and a type

C¹ algorithm (SMC, Acuros) as well as the difference between generic (SMCgen) and custom beam models (SMCcbm).^{7–9}

2 | Methods and Materials

2.1 | Patient data, treatment planning, and TPS dose calculation

A total of 100 patients were drawn consecutively from the clinical database, comprised of 20 case classes with five patients per case class (H&N, lung, lung SBRT/IMRT, lung SBRT/static-conformal, esophagus, breast, breast+SIB, breast supraclav+DIBH, spine met, brain SRT, rectum, cervix, cervix+pelvicLN, prostate+pelvicLN, brain, sinonasal, sarcoma arm, sarcoma leg, prosthesis, and palliative). These cases represent a broad range of patient groups treated in our department, with different treatment techniques and levels of challenges posed by treatment volumes and locations. For all patients except the lung and breast supraclav+DIBH cases, a free-breathing CT scan was acquired. For lung cancer patients, the mid-ventilation phase of a four-dimensional (4D) CT scan was used for delineation and dose calculation. For breast supraclav+DIBH, a CT scan performed in deep inspiration breath hold was utilized. A slice thickness of 3 mm was used for all cases except for H&N (2 mm), spine mets (1.5 mm), and brain SRT (1.5 mm).

The patients were treated using Varian TrueBeam linacs (Varian Medical Systems, Palo Alto, CA, USA) equipped with a Millenium or MilleniumHD 120 multi-leaf collimator (MLC). The beam qualities were 6 MV, 6 MV FFF, and 15 MV. All treatment plans were created in the Eclipse TPS (Varian Medical Systems, Palo Alto, CA, USA) version 15.6 using the Acuros XB algorithm for final dose calculation. Acuros XB is also a type C algorithm. The commissioning of Varian accelerators in the TPS was described in Refs. 5,10 An in-slice grid size of 2.5 mm was used for the calculations except for spine mets and brain SRT, where the grid size was 1.5 mm. Dose was reported as dose-to-medium. For each case class, one of five different standard treatment techniques (static-conformal, IMRT, FFF-IMRT, FFF-SBRT, and VMAT) was used for treatment planning. For all IMRT and VMAT plans, portal dose measurements using the on-board imager were performed as part of the patient-specific QA. The acceptance criterion was 95% of points fulfilling $\gamma < 1$, with a dose difference (DD) $< 3\%$ and distance-to-agreement (DTA) < 2 mm. All plans fulfilled this criterion.

¹ Type B algorithms are typically analytical and consider density variations by corrections to pre-computed electron kernels in water. Type C algorithms employ explicit electron transport in-situ and are typically either Monte Carlo methods or Linear Boltzmann equation solvers.

2.2 | Mobius3D

Dose calculation in M3D is based on a convolution/superposition algorithm using beam models that are pre-commissioned with standard reference datasets for the different beam energies of Varian TrueBeam machines. M3D offers options to customize its stock reference beam models to the user's individual treatment machines.¹¹ We followed the vendor's recommended procedure and confined the adaptation of the beam models to dose–monitor-unit calibration and Dosimetric Leaf Gap (DLG) correction.^{4,12} An isotropic voxel grid with a grid size of 3 mm was used for M3D calculations, except for H&N with an isotropic voxel size of 2.5 mm, and spine mets and brain SRT, with an isotropic voxel size of 2.0 mm. The voxel sizes were chosen automatically by M3D. The dose-to-material calibration cannot be set.

2.3 | SciMoCa

Dose calculation in SMC is based on a Monte Carlo algorithm.^{3,5,6} The vendor offers pre-commissioned beam models that were derived from pooled data (generic beam models) and fully commissioned, customized beam models that are adapted to individual treatment machines by the vendor.^{3,5,6} Customized beam models for the three beam modalities 6 MV, 6 MV FFF, and 15 MV were commissioned on the basis of 14 depth dose curves ($10 \times 10 \text{ mm}^2$ – $400 \times 400 \text{ mm}^2$), output factors, and cross-profiles of the maximum field size. These measurements were largely identical to the base data of the TPS, apart from some small field measurements that were not required for TPS commissioning. The physical leaf positions were calibrated for both beam model types to the same value from a set of standard DLG measurements used to commission the TPS. An isotropic voxel size of 2.5 mm was used except for spine mets and SRT cases, where the voxel size was 1.5 mm. The statistical uncertainty was chosen such that the volume receiving at least 70% of the maximum dose had an uncertainty $< 0.5\%$ for the total treatment plan dose. Dose was reported as dose-to-medium.

2.4 | Dose comparison

The 3D dose distributions from M3D and SMC were compared to the respective Acuros XB dose distributions using a global gamma evaluation¹³ with exclusion of doses below 5% of the maximum dose. The percentage of points fulfilling gamma < 1 , that is, the gamma pass rate (GPR), was denoted as GPR (DD, DTA), where a number of different DD and DTA settings were used: GPR (5%, 3 mm), GPR (3%, 3 mm), GPR (2%, 2 mm),

and GPR (2%, 1 mm). Furthermore, the difference in mean target dose (ΔD) versus Acuros XB was calculated. Dose–volume metrics were not investigated, since the 20 case classes varied so much in their treatment goals that a meaningful common parameter was not available.

Comparison of M3D versus Acuros XB was performed in the Mobius3D software, while the SciMoCa-Compare module of the SciMoCa software (Radiologica LLC, St. Louis, MO, USA) was used for comparisons of SMC. Since M3D does not support dose export, it was not possible to compute GPR of M3D and SMC by the same algorithm.

2.5 | Determination of action level

For the criteria ΔD and the four GPRs, action levels were determined that resulted in approximately $q = 0.02, 0.05, 0.1$, and 0.2 plan failure rate for each QA option by computing the percentiles and rounding to the next multiple of 0.5 and 5, respectively. Since GPR can be insensitive to target dose deviation, the criteria ΔD and GPR at $q = 0.1$ and 0.2 failure rate were also combined, that is, a plan fails if it fails according to ΔD AND GPR. Failing plans for either criterion were investigated for common causes.

The report of AAPM TG 218¹⁴ gives an alternative recommendation for establishing the action interval $2A$ from process observations (Equation 3):

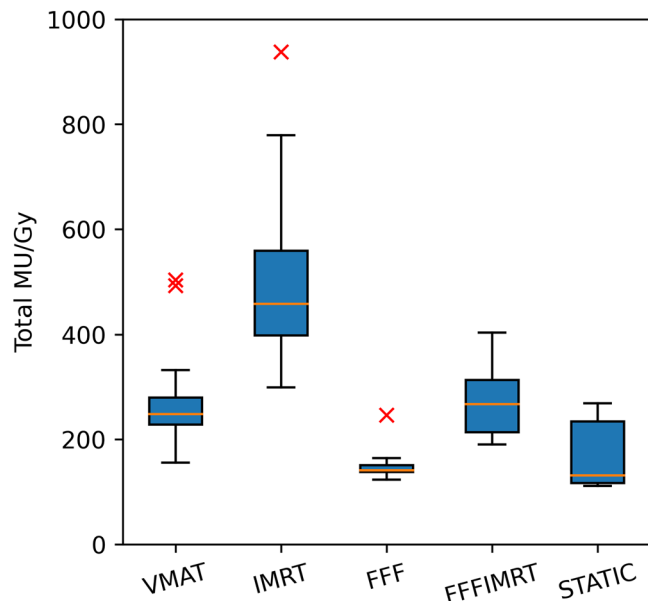
$$2A = \beta \sqrt{\sigma^2 + (\bar{x} - T)^2},$$

where T is the QA target value (e.g., GPR = 100), and σ^2 and \bar{x} are the observed variance and mean of the QA results. For the sake of clarity, the action levels would be $T - A$ and $T + A$ for a two-sided error and $T - A$ for a one-sided error like GPR. We can think of the root as the nominal variability of the QA process in the presence of systematic and random errors. In case the systematic component (from both ISDC and TPS) becomes very small, the root converges to the standard deviation of the QA process. Then, A becomes a $\beta/2$ -fold multiple of this standard deviation. Assuming a Gaussian distribution of x , q , and $\beta/2$ can be related via the normal distribution function, see Table 1. Notice that since the distribution of GPR values is truncated at 100, these values differ from those of a two-sided distribution like that of ΔD . Notice further that in the limit of vanishing systematic errors, the (tentative) TG 218 recommendation of $\beta = 6$ amounts to the stipulation that a QA failure is a 3-sigma event. To investigate the assumption of normal error distribution within the TG 218 formalism, we compared the nominal and observed failure rates for various action levels.

TABLE 1 Relationship between QA failure rates q and factors $\beta/2$ according to the TG 218 formalism for the action level and assuming a normally distributed error of the QA process.

q	0.0014	0.0027	0.02	0.023	0.046	0.05	0.1	0.2
$\beta/2$ for GPR	3	–	2.05	2	–	1.65	1.28	0.84
$\beta/2$ for ΔD	–	3	2.33	–	2	1.96	1.65	1.28

Note: Since the distribution of GPR values is right truncated, while the distribution of difference of the mean dose ΔD is two-sided, the values of $\beta/2$ differ.

**FIGURE 1** Modified box-and-whisker plot of the monitor unit expenditure (in terms of total MU/Gy) by treatment technique. Red line: median. Extents of box: central 50% of data points = interquartile range. Whiskers: min/max values within 1.5-fold interquartile range. Red crosses: outliers.

3 | Results

Figure 1 shows an overview of the monitor unit expenditure for the five treatment techniques. Of the 20 case classes, 2 were treated with 6FFF stereotactic conformal beams (lung-SBRT/static-conformal and brain SRT), 1 was treated with 6FFF IMRT (lung-SBRT/IMRT), 4 were treated with static conformal fields (breast, breast-supraclav+DIBH, breast-SIB, and palliative), 4 were treated with static beam IMRT (lung, arm, leg, and esophagus), and the remaining 9 with VMAT.

The GPR results for the four DD/DTA criteria combinations are shown in Figure 2.

The results of the difference of mean target dose ΔD are displayed in Figure 3. The mean $\Delta D \pm 1$ SD across all case classes was $1.75\% \pm 2.41\%$ for M3D, $-0.17\% \pm 1.03\%$ for SMCgen, and $-0.11\% \pm 0.78\%$ for SMCcbm.

From these findings, the action levels as a function of QA failure rate q were determined. For the four different gamma criteria, the results are shown in Figure 4. The action levels for the absolute difference in mean target dose $|\Delta D|$ are shown in Figure 5.

For the analysis of the effectivity of the combined application of both evaluation criteria, GPR and absolute difference in mean target dose $|\Delta D|$, action levels 1 and 2 were determined for $q = 0.1$ according to the conversion factors of Table 1 and rounded to the next multiple of 5 and 0.5, respectively (see Table 2). Action level 3 is a further restriction over action level 2, being determined for $q = 0.2$, and is added for better analysis of the difference between generic and custom beam models.

The distributions for the combined criteria with respect to these action levels are shown in Figure 6 and detailed in Table 3. In the same table, we show the failure rates that would be expected from the TG 218 formalism for these action levels, assuming a Gaussian distribution of QA results as in Table 1.

4 | DISCUSSION

This investigation aimed to quantify the benefit of accuracy for different ISDC concepts in terms of sensitivity at a constant workload, that is, a failure rate. Since the measurement of sensitivity requires a set of test cases with known or stipulated QA status, it always contains a degree of arbitrariness. We chose our approach because it affords greater independence from the set of test cases and relies only on the assumption that tighter QA criteria would in any case increase sensitivity to less than obvious errors—at a controlled failure rate. In addition, our method does not rely on assumptions about the distribution of QA results but determines action levels solely from percentiles of the results for a large number of cases, that is, by counting. This makes the method robust but perhaps too cumbersome for general usage.

An alternative approach to arrive at action levels was presented in the report of AAPM TG 218.¹⁴ It requires a smaller sample size but invokes assumptions about the distribution of QA results. The action levels depend on the choice of a parameter β , which is somewhat elusive. However, under certain assumptions, β can be related to the failure rate q via a Gaussian distribution function. We tested this hypothesis because the TG 218 approach could provide a shorter route to institution-specific action levels than ours, potentially requiring fewer cases.

We found that for QA criteria that are two-sided (ΔD), the failure rate, as expected from the action level, matches the observed failure rate in 6 out of 6 cases within uncertainties (Table 3). The agreement was much

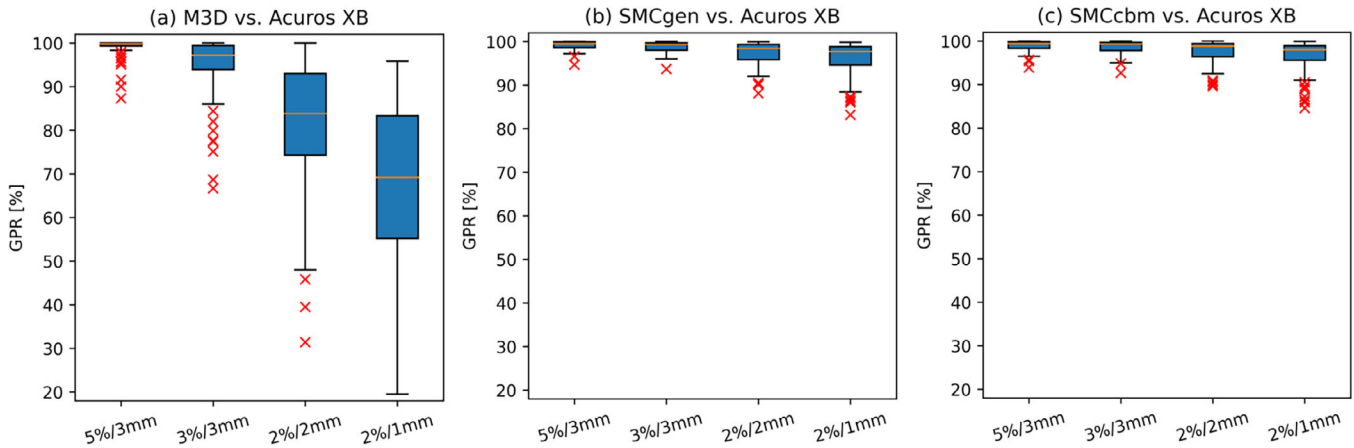


FIGURE 2 Modified box-and-whisker plots of the gamma pass rates GPR (DD, DTA) for different combinations of dose difference (DD) and distance-to-agreement (DTA) for (a) M3D, (b) SMCgen, and (c) SMCcbm versus Acuros XB, respectively. Red line: median. Extents of box: central 50% of data points = interquartile range. Whiskers: min/max values within 1.5-fold interquartile range. Red crosses: outliers.

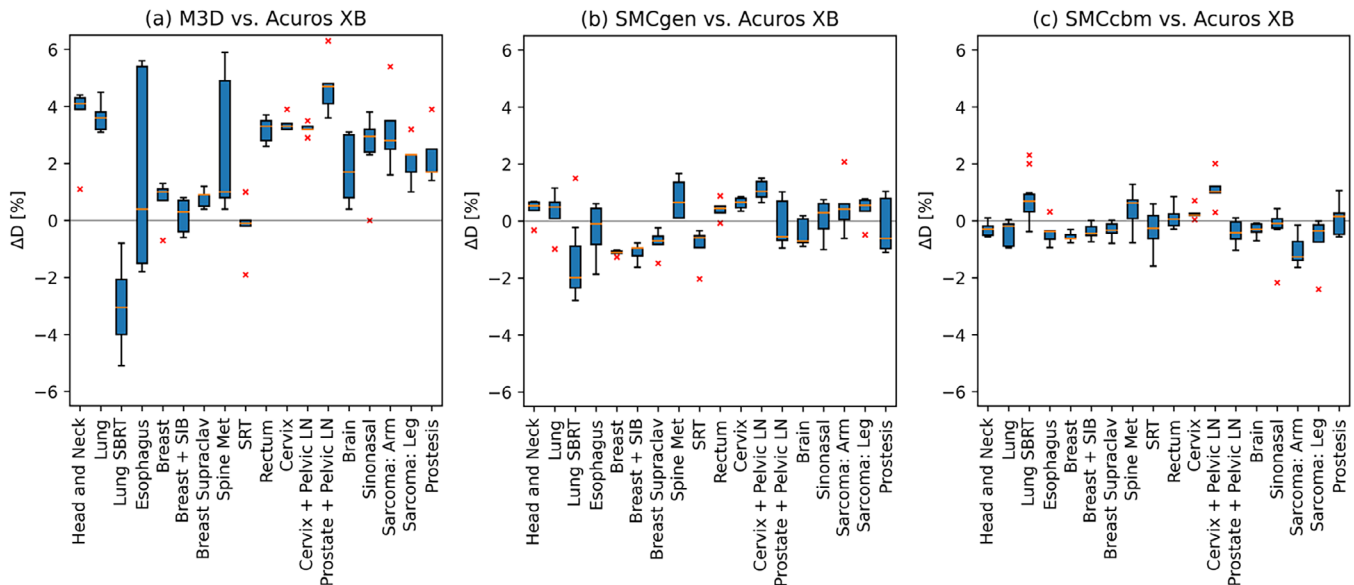


FIGURE 3 Modified box-and-whisker plots of the difference of mean target dose ΔD for (a) M3D, (b) SMCgen, and (c) SMCcbm versus Acuros XB, differentiated per case class. Positive values indicate the ISDC dose is higher than the TPS dose. Red line: median. Extents of box: central 50% of data points = interquartile range. Whiskers: min/max values within 1.5-fold interquartile range. Red crosses: outliers.

worse (1 out of 6) for the one-sided GPR criterion, particularly when the mean GPR was close to 100%. In this case, the error distribution becomes very asymmetrical with a heavy tail and is therefore not well approximated by a Gaussian. For a distribution of GPR like those for SMC with mean GPR \pm standard deviation between $96\% \pm 2\%$ and $98\% \pm 1.4\%$ it seems to be possible to correct $\beta/2$ by a factor 1.45, which is likely influenced by the mix of test cases. A more thorough investigation is beyond the scope of this manuscript.

Irrespective of the approach, both methods establish a link between action level and failure rate for a particular combination of ISDC and TPS (each with its own

particular quality of commissioning). Hence, the pair of action level/failure rate can serve as an inter-institutional measure of the overall quality of treatment planning and its QA.

In our investigation, the gamma criterion applied to the entire extent of the dose distribution and therefore lacks specificity, especially for small target volumes. This is the reason why we also considered a target-centric parameter, the difference between the mean target dose of the ISDC and the TPS in parallel with gamma. Also, the GPR is affected by Monte Carlo statistical noise,¹⁵ while the mean dose is the most robust parameter against such noise. It has been shown³ that the effect

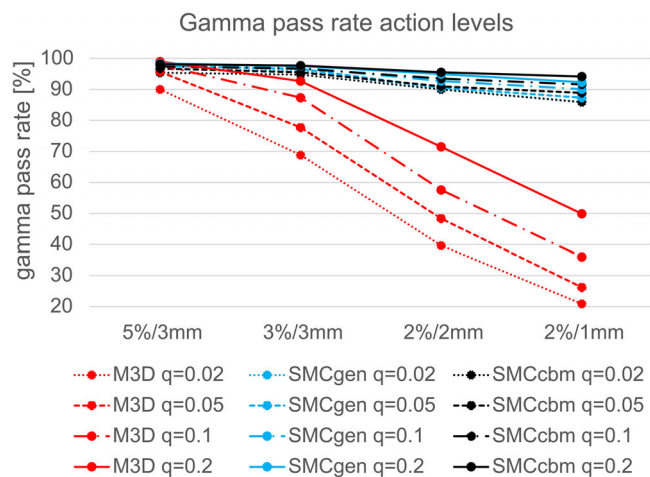


FIGURE 4 Action level in terms of gamma pass rate for each of the four gamma criteria as a function of test failure rate q for M3D (red), SMCgen (blue), and (c) SMCcbm (black) versus Acuros XB. Failure rates were $q = 0.02$ (dotted), $q = 0.05$ (dashed), $q = 0.1$ (dot-dashed), and $q = 0.2$ (solid).

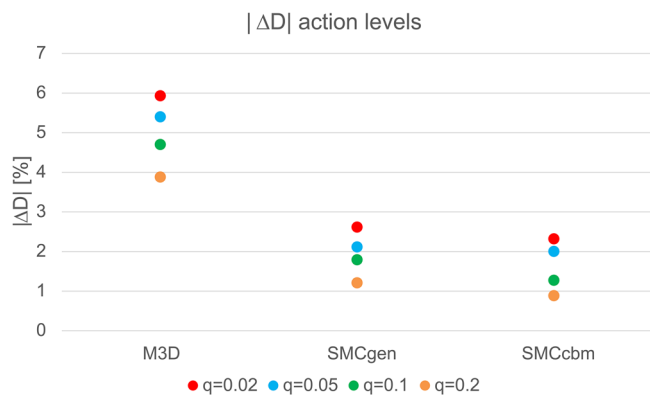


FIGURE 5 Action level in terms of the absolute difference in mean target dose $|\Delta D|$ depending on test failure rate q for M3D (left), SMCgen (center), and (c) SMCcbm (right) versus Acuros XB. Failure rates were $q = 0.02$ (red), $q = 0.05$ (blue), $q = 0.1$ (green), and $q = 0.2$ (orange).

of a 0.5% statistical uncertainty on GPR with a small DTA is rather small, in the order of $< 1.5\%$ variation of GPR, which is the same magnitude as the variability of different implementations of the method.¹⁵ Naturally, more case class specific criteria like region-of-interest restricted gamma analysis or dose-volume parameters would be used in practice but were not at our disposal here as they would not allow us to view many different case classes simultaneously.

The central results of our study are given in Table 3 and Figure 6. At its ideal action level of GPR (3,3) $\geq 90\%$, $|\Delta D| \leq 4.5\%$, M3D fails in both criteria in 3% of the cases. These cases were highly modulated IMRT treatments (two esophagus and one arm sarcoma). In terms of $|\Delta D|$ patterns of failure, there is a clear finding that lung SBRT cases, treated with 6FFF-IMRT beams, consistently scored a lower dose by M3D; a finding which

TABLE 2 Action levels AL1, AL2 for $q = 0.1$ failure rate and AL3 for $q = 0.2$ failure rate for the combined criteria analysis.

Action level (derived from ISDC)	1 (M3D)	2 (SMCcbm)	3 (SMCcbm)
$ \Delta D $	$\leq 4.5\%$	$\leq 1.5\%$	$\leq 1.0\%$
Gamma	GPR (3,3) $\geq 90\%$	GPR (2,2) $\geq 95\%$	GPR (2,1) $\geq 95\%$

has been reported previously.¹⁶ With respect to gamma failures, there seems to be a tendency towards lower GPR for arm sarcoma cases treated again with 6MV-IMRT. With the more stringent action level 2, the failure rates rise substantially, with 56% failing both criteria. There seems to be no apparent patterns or clusters of cases for outliers.

Several studies that evaluated the agreement of M3D with TPS and measurements also resulted in GPR (3,3) around 90%–95%,^{17,18,19} which is in keeping with our results. The bias towards positive ΔD in Figure 3a suggests that the DLG value is too high. However, the static beam cases also present with a higher average mean DD, the spread between the highest and lowest ΔD is rather symmetrical, and the deviation is driven primarily by VMAT cases, even though IMRT cases have a higher monitor unit expenditure (Figure 1). Difficulties with finding a general DLG value, that is, optimum for both VMAT and IMRT have been reported.²⁰ All of this suggests that the deviation of the mean ΔD from 0 is multi-factorial.²¹

At the action level 2 of GPR (2,2) $\geq 95\%$ and $|\Delta D| \leq 1.5\%$, the Monte Carlo ISDC fails in both criteria in 1% (generic beam model) and 2% (custom beam model) of cases. In the single-criteria and both-criteria pass rates, the custom beam models score somewhat better than the generic models (Table 3 and Figure 3b,c).

Both SMC computations have in common that breast and arm sarcoma cases present with lower GPR. This can, in fact, be traced back to a difference in the patient density model between Acuros XB and SMC. Figure 7 gives three examples that show the location of the low GPR volumes. On the left is an arm sarcoma case that illustrates a difference in the handling of the patient outline, which contains more density-filled voxels in SMC than in Eclipse. This is due to a different interpretation of partially filled voxels at the patient outline. The middle image is a leg sarcoma case with a dummy contour extending into the space between the fixation device and the leg. The difference in dose computation is limited to the dummy contour, in which SMC computes dose in air while Eclipse does not. Finally, the right image shows a breast case where the immobilization device and patient support structures are included in the density model but have a different material composition in SMC versus Eclipse.⁵ To a large extent, the drop in GPR could be avoided if the gamma distribution was not evaluated over the entire dose calculation volume. This also underlines the complementary nature of the comprehensive,

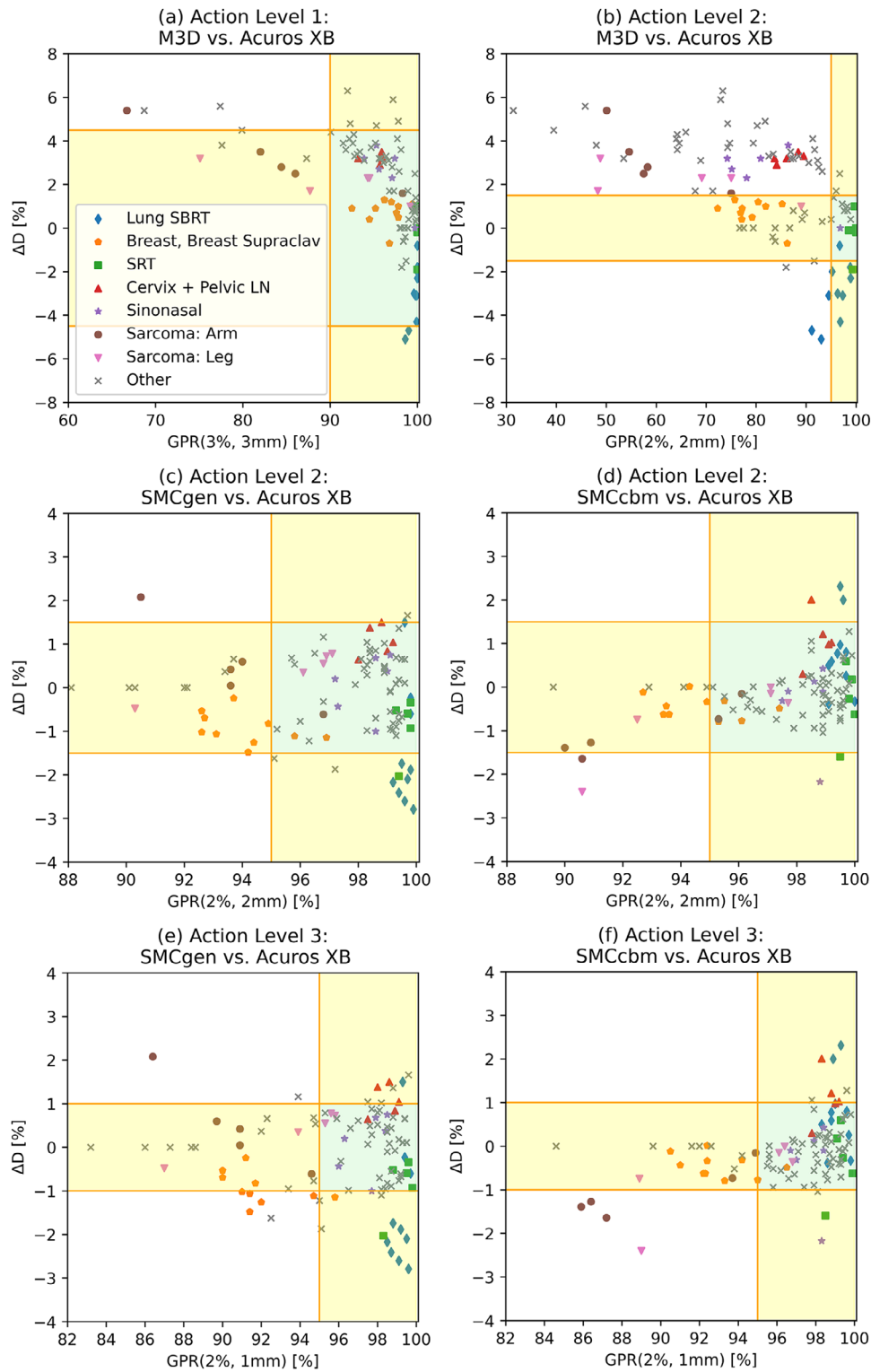


FIGURE 6 Scatter plots of combined criteria scores of each case for M3D with action level 1 (a), M3D with action level 2 (b), SMCgen with action level 2 (c) and action level 3 (e), and SMCcbm with action level 2 (d) and action level 3 (f). The green zone denotes the area where both criteria, gamma pass rate and absolute difference in mean target dose $|\Delta D|$, are fulfilled. The yellow zones denote the areas where only one criterion is fulfilled. White zones denote the areas where none is fulfilled.

TABLE 3 Pass/Fail rates for the combined application of both criteria, gamma pass rate and absolute difference in mean target dose $|\Delta D|$, for M3D with action level 1 and action level 2, and SMCgen and SMCcbm with action level 2 and action level 3.

ISDC action level	M3D 1	M3D 2	SMCgen 2	SMCcbm 2	SMCgen 3	SMCcbm 3
Passed both	82%	13%	70%	80%	53%	69%
Failed both	3%	56%	1%	2%	8%	4%
Failed gamma	11%	78%	19%	15%	29%	22%
Failed $ \Delta D $	10%	65%	12%	7%	26%	13%
Expected fail rate from AL gamma	$11 \pm 3.3\%$	$42 \pm 6.5\%$	$10 \pm 3.2\%$	$8 \pm 2.8\%$	$18 \pm 4.2\%$	$14 \pm 3.7\%$
Expected fail rate from AL $ \Delta D $	$13 \pm 3.6\%$	$62 \pm 7.9\%$	$13 \pm 3.6\%$	$5 \pm 2.2\%$	$31 \pm 5.6\%$	$17 \pm 4.1\%$

Note: The two bottom rows give the predicted failure rate via the TG 218 formalism, assuming a Gaussian distribution of QA results. Errors of expected failure rates are binomial estimates.

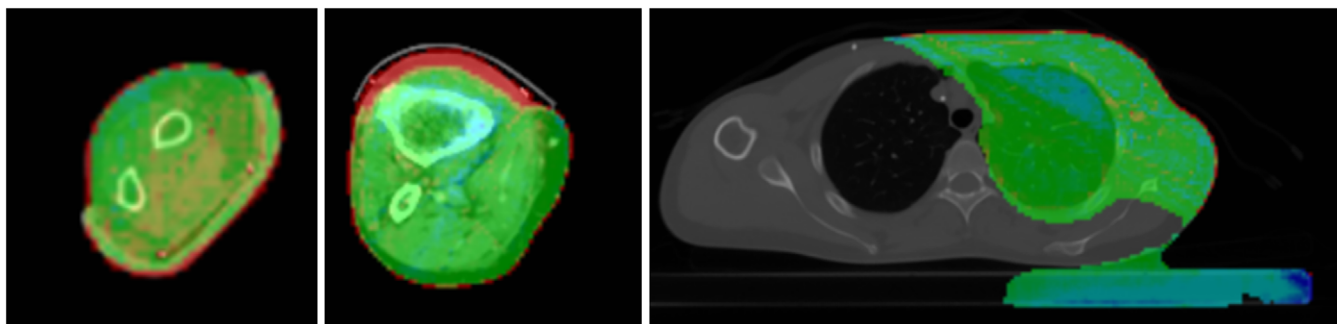


FIGURE 7 Volumes of low gamma pass rate (GPR) for three cases computed with SMC. The arm (left) and leg (center) sarcoma cases show a higher SMC dose at the patient outline (red area in the gamma plot). This stems from a different handling of the patient outline in the density models of Eclipse and SMC. In the middle image, only SMC computes dose in a dummy volume that extends outside the patient outline. The patient fixation of the breast case (right) shows a lower dose by SMC, which originates from a different translation of Hounsfield values into material properties by Eclipse and SMC (dose difference occurs in the foam core).⁵

spatially resolved, yet unspecific gamma analysis and the compounded, yet target-focused mean target dose.

A further case class with consistent discrepancies could be found for SMCgen, where all five lung SBRT-IMRT cases presented with a lower mean target dose than in Eclipse; this could not be observed for SMCcbm, that is, is caused by the beam model. In fact, these five cases were the only application of 6FFF-IMRT, so that this observation could be attributed to the generic beam model, which did not seem to be a good fit for the actual beam properties. In case such systematic discrepancies are observed, the issue may only be resolved by well-designed measurements specific to the hypothetical cause.

To investigate the performance of the generic beam models further, the SMC options were evaluated with the stricter action level 3, see Table 3. Here, the performance gap between the custom and the generic models widened compared to action level 2. The percentage of cases passing both criteria were 69% and 53% for custom and generic models, respectively. Even if the problematic 6FFF model had been removed from the analysis, a gap of 11% remained. For SMCcbm, only 4% of cases failed both criteria, while 8% failed both criteria for SMCgen. The two ISDC had only one failed

case (an arm sarcoma) in common, whose cause for failure could be traced back to the aforementioned differences in the patient model. Even though the overall performance of generic beam models was not markedly different from custom beam models for action level 2, the benefits of custom beam models begin to show at action level 3, where SMCgen presents twice as many QA fails than SMCcbm, or in other words: gratuitously doubles the workload. We conclude that it cannot be taken for granted that a generic model is always a sufficiently good match for a specific beam, as in case of the 6FFF model here. This obvious assertion also applies to the generic models of M3D, where there was no evidence that any one of the three models performed better than the others.

The observable difference between individually commissioned and generic beam models also gives an indication of the magnitude of commissioning errors of the TPS that may be picked up by SMC at action level 3. The criterion GPR (2,1) of action level 3 was also established as optimal in conjunction with a Monte Carlo TPS via a receiver-operator analysis,³ using custom beam models for Elekta accelerators. We interpret this as evidence that the typical accuracy of TPS with different flavors of a type C dose algorithm is indeed similar. It

was also found that the criterion GPR (2,1) can be satisfied by SMCcbm in comparison to measurements with detector arrays.^{6,22}

The majority of published validations of ISDC arrived at the conclusion that a GPR (3,3) > 90%–95% or GPR (3,2) > 90%–95% represents a satisfactory agreement to the TPS.^{17–19} The report of AAPM TG219² suggested the criterion of GPR (3,2) > 95%. These recommendations stem from a patient-centric, clinical view, that is, would avoid *random* serious mistreatments. The recent report of AAPM TG360²³ pointed out that TPS failure modes such as incorrect output factors or MLC modeling are dominating sources of errors of PSQA, which ties in with previous audit findings.²⁴ This suggests that *systematic* TPS inaccuracies play a significant yet overlooked role in PSQA. These systematic inaccuracies often originate from errors in the base data for commissioning the TPS and the ISDC. A recent paper by Kowatsch et al.²⁵ investigates the causes, severity and frequency of various error types present in the base data submitted for SMC custom beam modeling, emphasizing the benefit of independent commissioning. From our study, it appears that a GPR (2,1) > 95% is a suitable criterion for scanning for commissioning errors in either ISDC or TPS and should not result in failure rates much > 20%. While this failure rate may be acceptable for screening for systematic errors, it is clearly not acceptable for PSQA. In combination with other criteria, it may become more workable if we assume that a well-commissioned type C TPS (like the Acuros XB here or Monaco³) and a well-commissioned type C ISDC like SMCcbm operate at the same level of accuracy, action level 3 produces a background of 4% PSQA failures, of which 2% may be due to the TPS and 2% may be due to the ISDC. In contrast, at action level 2, M3D produces 27 times more PSQA failures than SMCcbm (Table 3).

It appears that the findings of TG360 call for an ISDC whose accuracy matches or exceeds that of the TPS to be surveyed. This is just another way of restating the recommendation of the ICRU83 report.¹ The report of TG360²³ also lends a different angle to the utilization of generic beam models for ISDC, which obviously require a thorough assessment of whether they match the specific beam properties, especially output factors and MLC modeling, for example, correct leaf position calibration.

Summarizing, we find that higher accuracy of an ISDC enables stricter acceptance criteria without increasing the workload. To capture also systematic errors of TPS commissioning in addition to random gross treatment plan errors, the action level needs to be adjusted to the capabilities of the TPS. For type C dose calculation TPS, a PSQA action level of GPR (2,1) > 95% OR $|\Delta D| < 1\%$ or stricter is feasible. Generic beam models require strict verification of their agreement with individual beam properties to guarantee this level of performance. Failure to achieve high pass rates in excess

of 95% at the above action level may indicate that the beam models of ISDC or TPS need attention.

5 | CONCLUSION

Monte Carlo dose calculation provides a significant benefit for ISDC as patient-specific quality assurance, allowing substantially more stringent acceptance criteria than an analytical algorithm. The performance of generic beam models can match that of customized beam models, but they still need to be validated against the individual beam properties. Apart from dose calculation, the generation of the patient density model is a source of substantial variability. This is one cause of low gamma pass rates, which could be ameliorated by restricting gamma analysis to the most relevant volumes. Depending on the quality of the commissioning of the TPS, a gamma (2%, 1 mm) pass rate > 95% is a feasible choice for Monte Carlo based ISDC with custom beam models.

AUTHOR CONTRIBUTIONS

Lone Hoffmann and Ditte Sloth Møller conceived the study and wrote the manuscript. Mai-Britt Linaa conducted data collection.

ACKNOWLEDGMENTS

M.L.A. and M.S. of sciRT consulted on the design of the study and preparation of the manuscript.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

ORCID

Lone Hoffmann 

<https://orcid.org/0000-0002-3575-0421>

REFERENCES

1. The International Commission on Radiation Units and Measurements. Report 83: prescribing, recording, and reporting photon-beam intensity-modulated radiation therapy (IMRT). *J ICRU*. 2010;10(1):1–106.
2. Zhu TC, Stathakis S, Clark JR, et al. Report of AAPM Task Group 219 on independent calculation-based dose/MU verification for IMRT. *Med Phys*. 2021;48:e808–e829.
3. Kowatsch M, Szeverinski P, Clemens P, Künzler T, Söhn M, Alber M. Sensitivity and specificity of Monte Carlo based independent secondary dose computation for detecting modulation-related dose errors in intensity modulated radiotherapy. *Z Med Phys*. 2023;S0939-3889(23):00117–4. doi:10.1016/j.zemedi.2023.10.001
4. Fontenot JD. Evaluation of a novel secondary check tool for intensity-modulated radiotherapy treatment planning. *J Appl Clin Med Phys*. 2014;15(5):207–215.
5. Hoffmann L, Alber M, Söhn M, Elstrøm UV. Validation of the Acuros XB dose calculation algorithm versus Monte Carlo for clinical treatment plans. *Med Phys*. 2018;45(8):3909–3915.
6. Adamson J, Erickson BG, Wang C, et al. Generation, validation, and benchmarking of a commercial independent

- Monte Carlo calculation beam model for multi-target SRS. *Z Med Phys.* 2023;S0939-3889(23):00093-4. doi:10.1016/j.zemedi.2023.08.004
7. Knöös T, Wieslander E, Cozzi L, et al. Comparison of dose calculation algorithms for treatment planning in external photon beam therapy for clinical situations. *Phys Med Biol.* 2006;51(22):5785–5807.
 8. Ojala JJ, Kapanen MK, Hyödynmaa SJ, Wigren TK, Pitkänen MA. Performance of dose calculation algorithms from three generations in lung SBRT: comparison with full Monte Carlo-based dose distributions. *J Appl Clin Med Phys.* 2014;15(2):4–18.
 9. De Martino F, Clemente S, Graeff C, Palma G, Cella L. Dose calculation algorithms for external radiation therapy: an overview for practitioners. *Appl Sci.* 2021;11:6806.
 10. Hoffmann L, Jørgensen MB, Muren LP, Petersen JB. Clinical validation of the Acuros XB photon dose calculation algorithm, a grid-based Boltzmann equation solver. *Acta Oncol.* 2012;51(3):376–385. doi:10.3109/0284186X.2011.629209
 11. Nelson CL, Mason BE, Robinson RC, Kisling KD, Dirsner SM. Commissioning results of an automated treatment planning verification system. *J Appl Clin Med Phys.* 2014;15(5):57–65.
 12. McDonald DG, Jacqmin DJ, Mart CJ, et al. Validation of a modern second-check dosimetry system using a novel verification phantom. *J Appl Clin Med Phys.* 2017;18(1):170–177.
 13. Low DA, Dempsey JF. Evaluation of the gamma dose distribution comparison method. *Med Phys.* 2003;30(9):2455–2464.
 14. Miften M, Olch A, Mihailidis D, et al. Tolerance limits and methodologies for IMRT measurement-based verification QA: recommendations of AAPM Task Group No. 218. *Med Phys.* 2018;45(4):e53–e83. doi:10.1002/mp.12810
 15. Clasie BM, Sharp GC, Seco J, Flanz JB, Kooy HM. Numerical solutions of the gamma-index in two and three dimensions. *Phys Med Biol.* 2012;57(21):6981–6997.
 16. Hillman Y, Kim J, Chetty I, Wen N. Refinement of MLC modeling improves commercial QA dosimetry system for SRS and SBRT patient-specific QA. *Med Phys.* 2018;45(4):1351–1359. doi:10.1002/mp.12808
 17. Clemente-Gutiérrez F, Pérez-Vara C. Dosimetric validation and clinical implementation of two 3D dose verification systems for quality assurance in volumetric-modulated arc therapy techniques. *J Appl Clin Med Phys.* 2015;16(2):198–217.
 18. Han C, Yi J, Zhu K, et al. Cross verification of independent dose recalculation, log files based, and phantom measurement-based pretreatment quality assurance for volumetric modulated arc therapy. *J Appl Clin Med Phys.* 2020;21(11):98–104.
 19. Li F, Price M. Commissioning of Mobius3D/FX for patient-specific quality assurance: the CUIMC-NewYork Presbyterian Hospital experience. *J Appl Clin Med Phys.* 2024;25(2):e14183. doi:10.1002/acm2.14183
 20. Zhao X, Baur M, Wall PDH, Laugeman E. Tuning a secondary dose verification software for a CT-guided online adaptive delivery system. *J Appl Clin Med Phys.* 2025;26(2):e14563. doi:10.1002/acm2.14563
 21. Kim J, Han MC, Lee E, et al. Detailed evaluation of Mobius3D dose calculation accuracy for volumetric-modulated arc therapy. *Phys Med.* 2020;74:125–132.
 22. Erickson B, Cui Y, Alber M, et al. Independent Monte Carlo dose calculation identifies single isocenter multi-target radiosurgery targets most likely to fail pre-treatment measurement. *J Appl Clin Med Phys.* 2024;25(6):e14290. doi:10.1002/acm2.14290
 23. O'Daniel J, Hernandez V, Clark C, et al. Which failures do patient-specific quality assurance systems need to catch? *Med Phys.* 2025;52(1):88–98.
 24. Glenn MC, Brooks F, Peterson CB, et al. Photon beam modeling variations predict errors in IMRT dosimetry audits. *Radiother Oncol.* 2022;166(1):8–14.
 25. Kowatsch M, Partoll E, Künzler T, et al. Quality assurance of beam base data in modern radiotherapy: a Monte Carlo simulation approach. *Biomed Phys Eng Express.* 2025;11(4). doi:10.1088/2057-1976/adde64

How to cite this article: Hoffmann L, Linaa M-B, Møller DS. Independent secondary dose calculation for patient-specific quality assurance: Quantitative benefit of Monte–Carlo and custom beam modeling. *J Appl Clin Med Phys.* 2025;26:e70265.
<https://doi.org/10.1002/acm2.70265>