Medical Physics
The International Journal of Medical Physics Research and Practice

Methods, software and datasets to verify DVH calculations against analytical values: Twenty years late(r)

Benjamin Nelms, Cassandra Stambaugh, Dylan Hunt, Brian Tonner, Geoffrey Zhang, and Vladimir Feygelman

Citation: Medical Physics **42**, 4435 (2015); doi: 10.1118/1.4923175
View online: http://dx.doi.org/10.1118/1.4923175
View Table of Contents: http://scitation.aip.org/content/aapm/journal/medphys/42/8?ver=pdfcov
Published by the American Association of Physicists in Medicine

Methods, software and datasets to verify DVH calculations against analytical values: Twenty years late(r)

Benjamin Nelms
Canis Lupus LLC, Merrimac, Wisconsin 53561

Cassandra Stambaugh
Department of Physics, University of South Florida, Tampa, Florida 33612

Dylan Hunt, Brian Tonner, Geoffrey Zhang, and Vladimir Feygelman[a)]
Department of Radiation Oncology, Moffitt Cancer Center, Tampa, Florida 33612

(Received 15 January 2015; revised 8 June 2015; accepted for publication 15 June 2015; published 2 July 2015)

Purpose: The authors designed data, methods, and metrics that can serve as a standard, independent of any software package, to evaluate dose-volume histogram (DVH) calculation accuracy and detect limitations. The authors use simple geometrical objects at different orientations combined with dose grids of varying spatial resolution with linear 1D dose gradients; when combined, ground truth DVH curves can be calculated analytically in closed form to serve as the absolute standards.
Methods: DICOM RT structure sets containing a small sphere, cylinder, and cone were created programmatically with axial plane spacing varying from 0.2 to 3 mm. Cylinders and cones were modeled in two different orientations with respect to the IEC 1217 *Y* axis. The contours were designed to stringently but methodically test voxelation methods required for DVH. Synthetic RT dose files were generated with 1D linear dose gradient and with grid resolution varying from 0.4 to 3 mm. Two commercial DVH algorithms—PINNACLE (Philips Radiation Oncology Systems) and PlanIQ (Sun Nuclear Corp.)—were tested against analytical values using custom, noncommercial analysis software. In Test 1, axial contour spacing was constant at 0.2 mm while dose grid resolution varied. In Tests 2 and 3, the dose grid resolution was matched to varying subsampled axial contours with spacing of 1, 2, and 3 mm, and difference analysis and metrics were employed: (1) histograms of the accuracy of various DVH parameters (total volume, D_{max}, D_{min}, and doses to % volume: $D99$, $D95$, $D5$, $D1$, $D0.03\ cm^3$) and (2) volume errors extracted along the DVH curves were generated and summarized in tabular and graphical forms.
Results: In Test 1, PINNACLE produced 52 deviations (15%) while PlanIQ produced 5 (1.5%). In Test 2, PINNACLE and PlanIQ differed from analytical by >3% in 93 (36%) and 18 (7%) times, respectively. Excluding D_{min} and D_{max} as least clinically relevant would result in 32 (15%) vs 5 (2%) scored deviations for PINNACLE vs PlanIQ in Test 1, while Test 2 would yield 53 (25%) vs 17 (8%). In Test 3, statistical analyses of volume errors extracted continuously along the curves show PINNACLE to have more errors and higher variability (relative to PlanIQ), primarily due to PINNACLE's lack of sufficient 3D grid supersampling. Another major driver for PINNACLE errors is an inconsistency in implementation of the "end-capping"; the additional volume resulting from expanding superior and inferior contours halfway to the next slice is included in the total volume calculation, but dose voxels in this expanded volume are excluded from the DVH. PlanIQ had fewer deviations, and most were associated with a rotated cylinder modeled by rectangular axial contours; for coarser axial spacing, the limited number of cross-sectional rectangles hinders the ability to render the true structure volume.
Conclusions: The method is applicable to any DVH-calculating software capable of importing DICOM RT structure set and dose objects (the authors' examples are available for download). It includes a collection of tests that probe the design of the DVH algorithm, measure its accuracy, and identify failure modes. Merits and applicability of each test are discussed. [http://dx.doi.org/10.1118/1.4923175]

Key words: dose volume histograms, quality assurance, software, external beam

1. INTRODUCTION

1.A. Background and rationale

The dose-volume histogram (DVH) is a useful data reduction construct that was developed in conjunction with three-dimensional (3D) treatment planning.[1–3] The DVH efficiently quantifies dose coverage for targets and sparing for organs-at-risk (OAR), reducing inherently 3D anatomy and dose data into a two-dimensional (2D) curve for each anatomical structure. The DVH has become a standard tool for evaluating radiation treatment plans and also serves as the basis for several common biological models.[2]

The AAPM TG-53 report on quality assurance for clinical radiotherapy treatment planning[4] included DVH accuracy as one of the outputs of the treatment planning system (TPS) that needs to be validated. However, this report provided little guidance on specific test methods, benchmark datasets, and acceptable (or expected) performance levels. Additionally, or perhaps consequently, medical device manufacturers have not published performance metrics in terms of the DVH accuracy of their systems nor are there standard and universal processes for clinical physicists to follow when commissioning this particular aspect of new systems. In the absence of specific guidance, the DVH calculations in clinical software systems can be considered undertested despite being universally used to evaluate and approve treatment plans.

1.B. Previous studies on the variation of DVH calculations

Others have performed studies on the variation in DVH methods and have noted that DVH calculation depends on multiple parameters, including but not limited to dose calculation grid resolution; discretization of a contour-based, anatomical region-of-interest (ROI) into a volume; and dose bin width of the histogram.[4–7] Varying methods can be programmed into software systems (i.e., TPS or third party plan review systems) to estimate DVH curves. For example, regular grid or random sampling techniques can be used to calculate DVHs.[8] As a result, DVH results reported by different systems for the same input dose grids and structures can vary.[8–10] To overcome these limitations for centralized multi-institutional plan analysis, different groups relied on in-house developed software packages to produce consistent DVH reporting.[8,9] While attempts were made to maximize numerical DVH calculation accuracy, such as using trilinear interpolation from the dose grid in SWAN software described by Ebert *et al.*,[8,11] the authors pointed out that it would not be correct to characterize their independent DVH calculation as "more accurate," as sufficient ambiguity in DVH reporting remained. Rather, in the context of the study, SWAN provided a reference to compare DVH calculations from different TPSs.[8,11] A similar approach was adopted by Straube *et al.*[12] but with synthetic test structures. Early on, Panitsa *et al.*[5] used simple geometrical structures for DVH quality control, but their standard based on the isodose distributions from the same TPS was fairly imprecise.

1.C. Purpose of this work

At our institution, there arose an immediate and practical impetus to perform a thorough analysis of DVH accuracy when a third party plan evaluation software system (PlanIQ, Sun Nuclear Corp., Melbourne, FL) was introduced in clinical practice. We use PlanIQ to process the outputs (plans, structures, dose, and CT images) from a number of various TPS. We noticed occasional examples where DVH parameters reported by the TPS differ somewhat from those recalculated by PlanIQ based on the TPS-generated DICOM RT dose and structure objects. This difference in DVH calculations using the same input data forced the question, "Which one, if any, is more accurate?"

To answer this question, we realized the need for specially designed structure sets and dose distributions with which ideal DVHs can be calculated analytically in closed form, independent of processing by any particular software system. The ideal analytical values could then serve as the ground truth when evaluating the outputs from any software system. This way, analyzing dose and volume *differences* is actually the examination of dose and volume *errors* caused by assumptions, limitations, or errors in the DVH calculation systems. While analytical DVH calculations for a single source were used as a reference in brachytherapy,[7,13] we are not aware of a similar approach reported for external beam DVH validation.

In this work, we supply analytical methods and fabricated datasets, and introduce a variety of comparison methods along with analysis software, with the aim to test DVH calculation accuracy of any system. We do proof-of-concept by testing the DVH curve results reported by two systems readily available to us: the PINNACLE TPS and the PlanIQ plan evaluation software. Though we test only two systems, our approach can be applied generally to test the DVH outputs of any system that is able to import DICOM RT structure set and RT dose objects. Therefore, these methods, datasets, and analytical comparison values can be applied by clinical physicists when commissioning a new TPS or plan review software system and medical device manufacturers during product validation.

2. METHODS

2.A. DVH calculations

2.A.1. General components of numerical DVH calculation

It is useful to break down the process of DVH calculation into basic components, each presenting a problem to be solved by a software algorithm. Understanding the multiple components facilitates the strategic design of test methods and datasets, the goals being to understand "if" there are errors or limitations and to deduce sources of error and failure mode(s). Understanding the multiple components also sheds light on why there is variation from one system to another, because each manufacturer must make design decisions on how to solve each problem. The general components of a DVH calculation system are summarized below.

— *3D digital rendering of 2D contour-based structures.* Anatomical structures, or ROIs, are described by lists of surface points in a closed planar loop—each one called a "contour"—defined on discrete axial slice positions [see Fig. 1(a)]. ROIs may be bifurcated. The coordinates for surface points for each ROI are stored in the DICOM RT structure set object. The coordinates in total make a series of 2D contours which must be rendered into a 3D volume, digitally represented as a collection of 3D pixels called "voxels." Ultimately, the purpose of the voxelated object is to answer a simple question about any

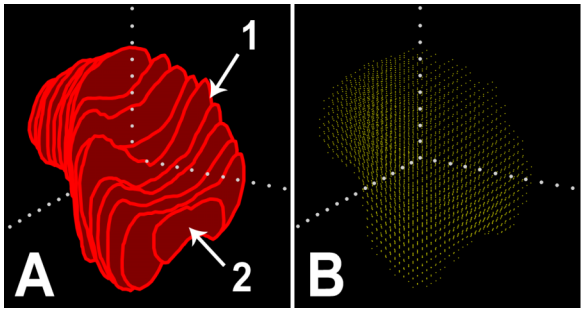

FIG. 1. (a) Series of 2D axial ROI contours plotted in 3D (view from anterior/superior/left). This example shows a prostate bed contoured on a CT series with 3 mm slice spacing. Region "1" highlights a representative region between slices where software must programmatically estimate the ROI volume boundary. Region "2" highlights the superior limit beyond which software must "end-cap" the volume. (b) Dose grid points calculated by the TPS at $2.5\times3.0\times2.5$ mm^3 grid resolution, plotted in 3D, and filtered to show only the points inside the ROI volume. In this example, for a 64.85 cm^3 ROI, 3486 discrete dose grid points fall within the volume. Supersampling by a factor of 3 in X, Y, and Z (and using interpolated dose values where new points do not fall on the original dose grid) would increase the number of dose points in the ROI to ~94 000.

generic point (X,Y,Z): Is the point inside or outside the ROI's volume? This question is elemental when trying to determine if a dose point (or dose voxel) of any location is to be "binned" as part of structure's DVH.

— *Structure end-capping and interslice interpolation.* When rendering 2D axial contours into high-resolution volumes, it is immediately apparent that the software must have solutions for how to "fill in" volumes in between slice locations (surface interpolation). A special consideration is how to treat axial contours at a superior or inferior edge of the ROI, which we will generally call end capping. Should the ROI's volume be truncated exactly at the last slice? Should it be extended halfway to the next, empty slice? Should it be extrapolated down to a point? Or should the user be able to configure the system in terms of end-capping method? Figure 1(a) highlights the regions in-between, and at the superior/inferior edges, of 2D contours.

— *Spatial resolution of input dose grid.* Independent of the RT structure set is the RT dose, which contains a 3D grid of point locations and dose values at each location. The resolution (or grid spacing, in three orthogonal directions) is set by the user in the TPS [see Fig. 1(b)]. The grid resolution must be sufficiently fine to capture steep dose gradients, peaks, and valleys. Dose at any point in between actual grid points must be interpolated from the input grid; coarse dose grid resolution will compromise accuracy of DVHs, isodose line rendering, and any other dose-derived graphic or metric.

— *DVH dose sampling spatial resolution.* When ROI contour shapes are either complex or small relative to the input dose grid resolution, dose supersampling will be required to produce an accurate DVH. That is, dose voxels binned for the DVH will need to be sufficiently small and comprised of voxel centers that do not coincide with the points in the input dose grid.

— *DVH dose sampling spatial alignment.* The alignment of the DVH dose sampling grid relative to the original voxelated ROI volume will have an impact and must be considered. For instance, if the DVH dose sampling grid is independent of the input dose grid and is not forced to have points that perfectly overlap the input dose grid, then local and global dose extremes (minimum and maximum) will be in error due to spatial "misses" and steep gradients will be degraded due to reliance only on interpolated point.

— *DVH dose bin width.* As dose voxels are processed and added to a ROI's histogram, each high-precision dose value must be indexed into a single DVH dose bin. The "bin width" (of the dose axis) is therefore of vital importance. Users must realize that, although presented as smooth curves, the cumulative DVHs are still histograms of discrete dose bins.

2.A.2. PINNACLE

PINNACLE v.9.8 was used. As previously reported,[8,14] for DVH calculations, PINNACLE uses regular sampling on the plan dose grid, with the user-defined dose bin size (≥1 cGy). The structure is automatically expanded from the boundary slices 1/2 CT slice width inferiorly and superiorly.[10] This was verified by computing the volumes of axially aligned cylinders with different axial contour spacings. Other than that, the details of the DVH calculation algorithm are not documented in public domain. However, armed with the precise analytical numbers, some aspects of this black box behavior can be deduced.

For PINNACLE, dose bin resolution was set to the finest possible increment of 1 cGy which resulted in approximately 2400 dose bins. This bin size did not exceed 0.4% of the lowest possible dose (2.5 Gy).

2.A.3. PlanIQ

A research version of PlanIQ (v2.1) was used. PlanIQ creates a 3D ROI voxel volume from axial slices first by generating a 3D surface then cutting into an orthogonal grid of planes at any size, resolution, and alignment. The resolution and alignment are forced to include the input dose grid points at their exact location. For small (as defined by the user) and/or complex ROI shapes, the ROI grid is supersampled (i.e., contains added planes in each of X, Y, and Z directions) in odd integer multiples (i.e., 3×, 5×, and 7×) until the ROI voxels number 40 000 or more. The choice to use odd supersample factors ensures that no slabs of "new" voxels (that will need to be interpolated) are centered exactly between the original dose plane locations. During structure voxelation, PlanIQ extends axial contours half-way to the next axial slice position in between slices as needed.

PlanIQ allows the user to define their desired treatment of end-capping, which can be set anywhere from zero (no end-capping, i.e., truncation at last slice) to up to half the distance to the next superior or inferior slice in the series of CT slice

locations. For these studies, PlanIQ was set with the latter, in order to mimic PINNACLE and to facilitate direct comparisons.

Dose voxels are guaranteed to be centered at the original dose grid locations with interpolated voxels in-between, when supersampled. Voxel elements are counted within each ROI volume and binned into equal DVH bin widths at whatever decimal precision (e.g., 0.0001 Gy) that results in 10 000 or more total dose bins from zero to the maximum dose in the grid. To ensure that even ROI surface points are included in the capture of the ROI's min and max dose, a postprocessing routine goes through each (X, Y, Z) coordinate comprising the ROI's contours, samples the interpolated dose at each location, and updates the ROI's min or max dose if applicable.

2.B. Synthetic structure sets and dose grids of variable resolution

Throughout this paper, the IEC 1217 coordinate system is used. Four synthetic CT datasets were produced with contiguous slices of 0.2, 1, 2, and 3 mm thickness. The structure sets consisted of single spheres, cylinders, and cones, inspired by work of Straube *et al.*[12] The structures in the DICOM RT format were generated programmatically in MATLAB (R2011a, MathWorks, Natick, MA), by finely discretizing analytical expressions for their axial cross sections (≤0.2 mm distance between the adjacent points). For cylinders and cones (Fig. 2), two orientations were used: with the major axis aligned along the Y-axis ("axial") and the Z-axis ("rotated").

The axial cross sections are circles of variable size for the sphere and axial cone, and circles of constant size for the cylinder. The axial cross sections are structure dependent for rotated shapes: a rectangle for the rotated cylinder and a hyperbolic cone section for the rotated cone (except for the triangle through the major axis). To generate axial coordinates, all structures were first finely sampled with 0.2 mm axial spacing. Geometrically, the structures just touched the first and last axial planes, resulting in cross sections which, strictly speaking, converged to either a point or a line. However, to avoid potential interpretation ambiguity, the intersection points or lines at the boundary axial slices were replaced with tiny polygons of negligible area. The structures were then subsampled with axial spacing of 1, 2, or 3 mm. The diameter of the sphere and the base circles of the cylinders and cones, as well as their height, was strategically set at 24 mm. Thus, we were able to keep the first and last slices in place, while subsampling was reduced to eliminating contours at the appropriate Y positions. The volumes were relatively small, 7.3, 11.0, and 3.7 cm^3 for the sphere, cylinder, and cone, respectively. The DICOM RT structure objects were imported into PlanIQ and PINNACLE using their respective standard DICOM interfaces.

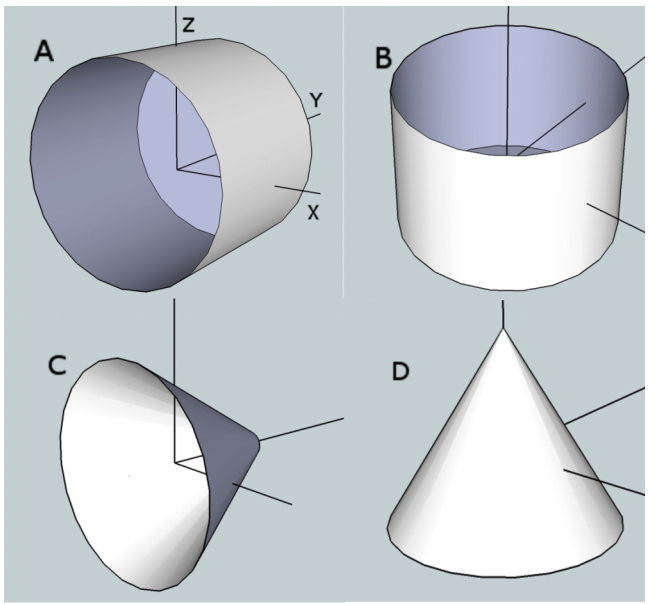

FIG. 2. Schematics of the structures in the IEC 1217 coordinate system. Axial (a) and rotated (b) cylinders. Axial (c) and Rotated (d) Cones.

Dose grids were also synthetic. First, an arbitrary dose grid of desired size and resolution was created in PINNACLE and exported as a DICOM RT dose object. That object was programmatically manipulated to replace the dose pixel values and then resaved as a DICOM RT dose file. Each dose grid had an exactly 16 Gy value at the center of the structure. One set of grids had a linear 1D dose gradient of 1 Gy/mm in the Y direction, and another in the X direction, resulting in the nominal D_{max} and D_{min} of 28 and 4 Gy, respectively. The simple linear dose gradient was used in order to facilitate the analytical calculation, which is somewhat limiting in that it is not necessarily representative of clinical gradients. The slope of 1 Gy/mm was chosen to provide ~6%/mm at the center of the structure. Naturally with the linear dose function, percentage dose error will be higher in the low dose and lower in the higher dose areas.

The smallest dose grid resolution was $0.4\times0.2\times0.4$ mm^3. Dose pixel dimensions below 0.4×0.4 mm^2 in the axial plane are not supported in PINNACLE. In addition, dose voxel cubes 1, 2, and 3 mm on a side were produced, corresponding to the set of structures' axial spacing.

For analysis with PlanIQ, these synthetic dose grids were imported via the standard DICOM interface. However, our PINNACLE license does not support import of RT dose objects, so a custom interface was built to convert DICOM RT dose into internal PINNACLE format that could be read in. It was verified that the dose fell in place of the original PINNACLE dose grid, which served as a starting point, without additional interpolation. At the end of these manipulations, we had a set of discretized structure and dose representations, with different resolutions, identical and properly aligned in PlanIQ and PINNACLE, and ready for DVH computation and comparison with the closed-form calculations.

2.C. Analytical DVH calculation

An analytical cumulative DVH formula, $V(D)$, can be derived for simple 3D geometrical objects. With dose changing in one direction, $D(x)$, for a given ROI, the infinitesimal volume, dV, may be written as the product of an area, A, and an infinitesimal length, dx, $dV = Adx$. The derivative of x with respect to dose, D, may be taken—assuming $D(x)$ is

invertible—and substituted in the infinitesimal equation for volume, $dV = A(dx/dD)dD$. The area, A, may itself be a function of x, in which case a substitution of $x = x(D)$ is necessary,

$$dV = A(x(D))\frac{dx}{dD}dD. \quad (1)$$

The dose-volume histogram function follows as the integral of Eq. (1) over appropriate limits,

$$V(D) = \int_{D}^{D_{\max}} A(x(D))\frac{dx}{dD}dD. \quad (2)$$

The volume in Eq. (2) can be integrated with respect to x, with appropriate limits, and then a substitution can be made to write x in terms of D,

$$V(D) = \int_{x(D)}^{x(D_{\max})} A(x)dx. \quad (3)$$

The two equations are identical; however, the integral in Eq. (3) may be easier to evaluate in closed form than the one in Eq. (2). Finally, the analytical DVH curve was modified to account for structure extension in both PINNACLE and PlanIQ by 1/2 axial spacing superiorly and inferiorly. This affects the maximum and minimum analytical dose when the dose gradient is along the Y-axis. In addition, the total volume is affected for the axial cone and cylinder. The effect of this expansion on the volume calculations for other structure/orientation combinations is negligible (<0.01%).

2.D. Test methods, metrics, and analysis software

Testing of DVH accuracy vs analytical (i.e., true) curves requires one to ponder different test methods, and for each appropriate metrics and acceptance criteria. Our strategy consisted of three test methods, each with a unique purpose and metric(s) befitting that purpose.

2.D.1. Test 1

In Test 1, the axial contour spacing was kept constant at 0.2 mm to essentially eliminate the variation and/or errors associated with rendering axial contours into volumes, and to focus solely on the effect of altering the dose grid resolution in various stages from fine ($0.4\times0.2\times0.4$ mm^3) to coarse ($3\times3\times3$ mm^3). Analytical results for the following parameters per structure were compared to both PlanIQ (with supersampling turned on: Ref. 20) and PINNACLE: total volume (V); mean, maximum, and minimum dose (D_{mean}, $D_{\max}$, $D_{\min}$); near-maximum ($D1$, dose covering 1% of the volume) and near-minimum ($D99$) doses; $D95$ and $D5$; and maximum dose to a small absolute (0.03 cm^3) volume ($D0.03$ cm^3). We were primarily interested in the high and low dose regions because with the linear dose gradient, they correspond to the structure boundary and this is where the deviations are expected to occur.

2.D.2. Test 2

In Test 2, each of three subsampled axial contour spacings (1, 2, and 3 mm) was paired with the dose grid resolution of the same size (emulating common practice). The resolution range of 1–3 mm (slice spacing and dose grid) covers the range used in practice for modern radiation therapy. The same analysis and metrics used for Test 1 were used for Test 2. We chose to group the results across all permutations because we wanted the statistical results to represent the accuracy results across a range of common data resolutions. More comprehensive metrics statistics per data resolution were saved for Test 3, which is designed to be a more "diagnostic" technique to identify failure modes if failures are evidenced by Tests 1 or 2.

2.D.3. Test 3

Test 3 features the same data conditions of Test 2 but with a more comprehensive analysis technique that used a dedicated software system developed for the purpose, and as a result, producing different metrics. In this test, a large sample of volume errors (%) was collected by extracting many points across the continuum of each DVH curve pair (numerical vs analytical) and binning into a volume (%) error histogram, per DVH and for each dataset permutation. Analysis of volume differences is important to do in addition to dose difference analyses, such as those described earlier. While dose differences, such as those used in Tests 1 and 2, are conceptualized as the distances between the curves along a horizontal line drawn through the volume (vertical) axis, volume differences are the distances between curves along a vertical line drawn through the dose (horizontal) axis.

2.D.4. Special analysis software

No commercial software systems have tools customized for statistical comparison of two continuous DVH curves. To facilitate and semiautomate the generation of comparison metrics, we developed a simple software system called "CurveCompare."[21] CurveCompare allows the import of text or tabular data exported by TPS and other systems that compute DVH data. Curves are plotted and the user can calculate dose differences and/or volume differences over the entire range of the curves. Difference curves and histograms of differences (along with basic statistics such as minimum, maximum, and median differences; mean; and standard deviation) are immediately displayed.

For this work, analytical curve raw data were imported into CurveCompare as the reference curves, and the comparison curve data were imported from DVH raw data exported by PINNACLE and by PlanIQ. PINNACLE and PlanIQ use slightly different export formats, but the data were directly copied to the clipboard and parsed into the tabular data by CurveCompare, thus making them universally useful for all DVH systems that can export DVH data to a delimited text file.

2.E. Data presentation

Performing the tests above resulted in a large number of data points that do not lend themselves to a standard statistical presentation (mean values, etc.) because each was designed for a specific purpose and therefore unique. Therefore, one of

the primary metrics presented for Tests 1 and 2 is the count of the numerical data points deviating by more than 3% from the analytical value for extracted dose-at-volume points, along with the corresponding range of observed percent differences. It is important to note that the percent differences here are all normalized based on the local, as opposed to maximum, dose and are thus amplified for lower dose values such as D_{min}, $D99$, and $D95$.

For Test 3, volume error histogram analyses were performed with CurveCompare to directly compare numerical (PINNACLE or PlanIQ) vs analytical raw data. Curves were rendered as volume (cm^3) vs dose (Gy) and sampled (301 samples per curve) to generate statistics of volume (cm^3) error, which were then normalized per dataset to give volume (%) error. The software allows sampling at any resolution; we chose ~300 points per curve as a number sufficient to sample both high and low curve gradients. Statistical results of volume errors (minimum, maximum, mean, and standard deviation) per dataset and per numerical system were recorded in tabular form. Several example DVH curves, overlapping numerical with analytical, along with one example volume error histogram, are plotted in separate figures.

The results are followed by a discussion of the trends and an explanation of the observed effects when possible. In addition, histograms of the data showing more variation ($D99$, $D95$, $D5$, and $D1$) are presented. An interested reader can download the entire results table from the supplementary material,[15] along with the DICOM RT structure and dose objects used to generate it.[15]

3. RESULTS

3.A. Test 1—Constant 0.2 mm axial contour spacing with variable dose grid resolution

Summary results are presented in Table I and Fig. 3. For this test, PINNACLE produced 52 data points with deviation >3%, out of possible 340 (15%), while PlanIQ produced five (1.5%). Volume deviations were counted only once since the structure volume is independent of the gradient direction.

With the finest axial contour spacing (0.2 mm), PINNACLE and PlanIQ calculate the volume of each structure within 2% and 1% of the analytical value, respectively. For both PINNACLE and PlanIQ, D_{min} is the same for each structure and dose grid resolution and varies only with the gradient direction. All PINNACLE values are 7.5% low for the Z-gradient and 2.6% high for the Y-gradient, while PlanIQ data show no deviation from the analytical values with one decimal place precision. The D_{max} values have similar absolute deviations but because of the higher denominator, the reported percent differences are much smaller.

For PINNACLE, $D99$ values are independent of the dose grid resolution. Only the cylinders (axial and rotated) have their $D99$ values always within 3% of the analytical. For the sphere, $D99$ is 1.1% and 4.2% high with the Z- and Y-gradients, respectively. For all other structures, $D99$ is 4.4%–7.5% high. PlanIQ $D99$ values depend somewhat on the dose grid resolution. All three points exceeding the 3% threshold belong to the rotated cylinder, varying from 3.8% to 5.2% for the dose grid resolutions ranging from 1 to 3 mm.

TABLE I. Results of Test 1. Dose grid resolution is varied while axial contour spacing is kept at 0.2 mm. Numbers of points (n) exceeding 3% difference (Δ) from analytical are presented along with the range of % Δ. Total number of structure/dose combinations is $N = 40$ (20 for V).

	Numerical–analytical			
	PINNACLE		PlanIQ	
Parameter	n, with $\Delta > 3\%$	Range of Δ (%)	n, with $\Delta > 3\%$	Range of Δ(%)
V	0	−2.0 to 1.9	0	−0.4 to 0.9
D_{min}	20	−7.5 to 2.6	0	0.0
D_{max}	0	−1.1 to 1.1	0	0.0
D_{mean}	0	−1.9 to 0.0	0	−0.1 to 0.7
$D99$	20	−1.4 to 7.5	3	−1.8 to 5.2
$D95$	12	−6.6 to 5.2	2	−0.8 to 3.9
$D5$	0	−1.8 to 1.0	0	−0.4 to 0.8
$D1$	0	−0.9 to 2.2	0	−0.4 to 0.4
$D0.03\ cm^3$	0	−0.9 to 1.3	0	−0.4 to 0.3

For $D95$ in PINNACLE, the majority of deviations greater than 3% are associated with cylinders: the values are 5.2% high for the axial cylinder with the Z-gradient and 6.6% low for the rotated one with the X-gradient. The remaining four instances of $D95$ being high by 3.6% are all associated with the rotated cone. For PlanIQ, the deviations of $D95$ of more than 3% are

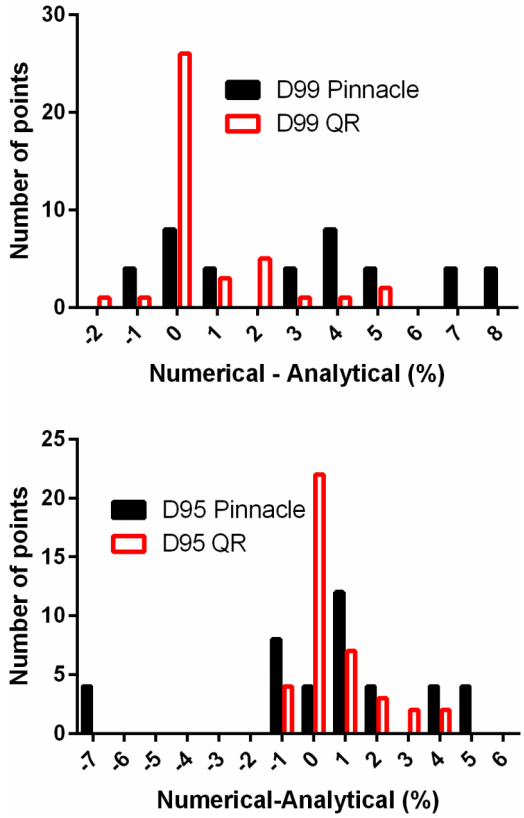

FIG. 3. Frequency distributions of percentage differences between numerical and analytical DVH parameters for PINNACLE and PlanIQ for Test 1 (varying dose grid resolution with constant axial contour spacing).

TABLE II. Results of Test 2. Dose grid resolution and contour spacing are matched at 1, 2, and 3 mm. Numbers of points (n) exceeding 3% difference (Δ) from analytical are presented along with the range of % Δ. Total number of structure/dose combinations for each parameter is $N = 30$ (15 for V).

	Numerical–analytical			
	PINNACLE		PlanIQ	
Parameter	n, $\Delta > 3\%$	Range of Δ (%)	n, $\Delta > 3\%$	Range of Δ (%)
V	0	−2.8 to 2.1	1	−4.2 to 0.6
D_{min}	30	−7.5 to 60	0	0.0
D_{max}	10	−5.1 to 1.1	0	0.0
D_{mean}	0	−1.9 to 0.0	0	0.0
$D99$	18	−4.2 to 44.4	11	−4.2 to 22.3
$D95$	19	−7.8 to 19.5	4	−2.9 to 7
$D5$	2	−3.6 to 5.3	0	−1.7 to 0.5
$D1$	7	−8.1 to 2.6	1	−3.9 to 0.8
$D0.03\,cm^3$	7	−7.8 to 0.9	1	−4.6 to 0.9

observed only for the rotated cylinder, with the Z-gradient and dose grid resolutions of 1 and 3 mm (3.9% and 3.6%, respectively). However, with the dose voxels of $0.4\times0.2\times0.4$ and $2\times2\times2$ mm^3, the difference is at least 2.6%, indicating roughly the same pattern.

3.B. Test 2—Variable contour spacing with variable (matched) dose grid resolution and discrete dose difference metrics

Overall in Test 2, out of possible 255 data points, PINNACLE exhibited deviation >3% in 93 points (36%) while 18 (7%) deviations were noted with PlanIQ. The results are summarized in Table II and Fig. 4. Between the two systems, all total structure volumes (cm^3) were within 3% of analytical except one. The PlanIQ number for the rotated cylinder with 3 mm spacing was 4.2% below nominal. The direction of this volume error is consistent with the fact that the rotated cylinder is constructed of the rectangles on the axial cross sections. As the distance between contours increases, the reconstructed polygon base in the XY plane deviates further from the ideal circle circumscribing it, resulting in the diminishing volume. The difference increases from −1.0% to −4.2% as the spacing changes from 1 to 3 mm.

For PlanIQ, the observed deviation is 0.0% for both D_{min} and D_{max}, while D_{mean} is 0.7% off for the rotated cylinder. PlanIQ's accuracy of D_{min} and D_{max} calculations is due to the special postprocessing step of PlanIQ that accounts for all dose values at surface points defined by the contour coordinates and not limiting the calculation to just the supersampled dose voxels contained by the volume (which by definition will be inside the surface points).

With PINNACLE, the deviation in those values depends on the gradient direction only. The minimum dose D_{min} is always 7.5% low for the Z-gradient. For the Y-gradient, the error is contour spacing-dependent, rising from 14.3% to 60.0% between 1 and 3 mm for all structures. This effect demonstrates that PINNACLE does not take the end-capping into consideration when reporting the minimum dose. A similar effect but with the opposite sign is seen for PINNACLE D_{max}. Because of the higher denominator, the deviation above 3% does not surface until the contour spacing reaches 2 mm, and

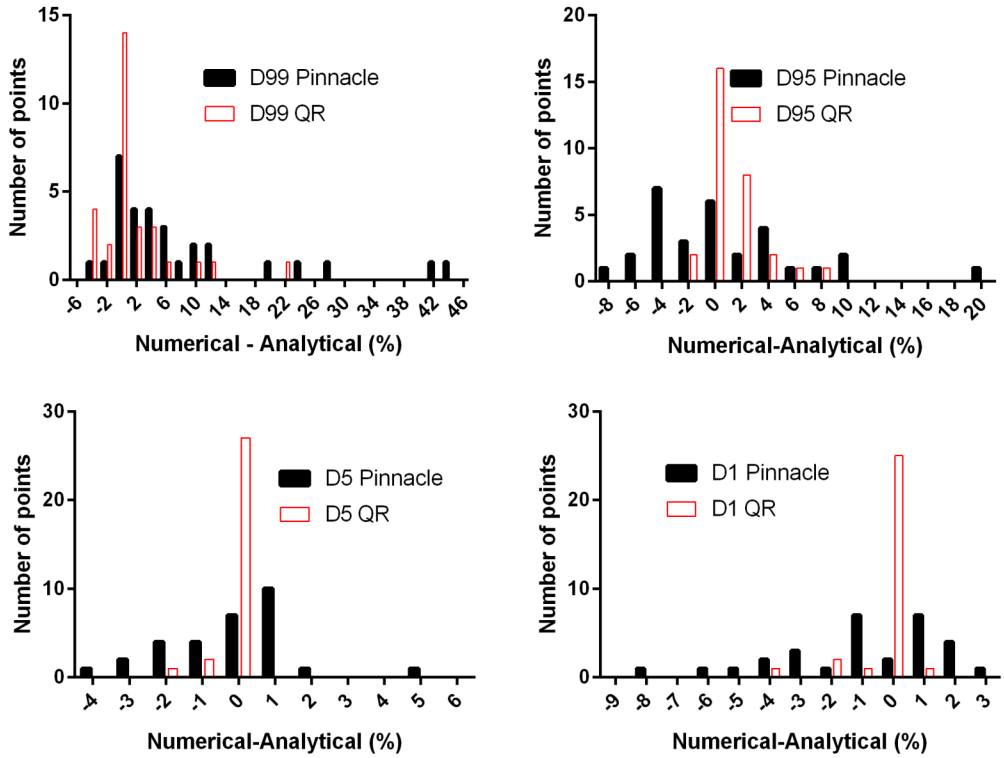

FIG. 4. Frequency distributions of percentage differences between numerical and analytical DVH parameters for PINNACLE and PlanIQ for Test 2 (matched dose grid resolution and axial contour spacing, varying from 1 to 3 mm).

then it increases from −3.4% to −5.1% as the spacing increase to 3 mm.

Similarly, PINNACLE $D99$ values almost always exceed the nominal for the Y-gradient, from 4.3% to 41.4%. The only exception is the rotated cone at 2 mm contour spacing (−3.5%). With PlanIQ, there is no clear pattern for the $D99$ deviations. With 1 mm contour spacing, only rotated cylinder shows differences above 3% (3.8% and 4.6% for the Z- and Y-gradients, respectively). For the 2 and 3 mm spacings, the largest errors are for the rotated cylinder with the Y gradient (11.5% and 22.3%, respectively), and for the sphere at 3 mm wither same gradient (9.8%). The remaining deviations scored in Table II range from −4.2% to 5.2%, with both extremes occurring with the 3 mm spacing.

With PINNACLE, 19 scored variations in $D95$ vary rather randomly and are observed with all three contour spacing values. The worst case scenario is the rotated cone with the Y gradient and 3 mm spacing (19.5%). For PlanIQ, all four scored deviations are associated with the rotated cylinder: one with 1 mm (Z-gradient), one with 2 mm (Y-gradient), and two with 3 mm contour spacing (both Z and Y-gradients). The values range from 3.5% to 7%. The Y-gradient error increases monotonically with the contour spacing, while the Z-gradient one dips below 3% for the 2 mm spacing.

Similarly to D_{max}, the largest variations at the high end of the PINNACLE DVH ($D5$, $D1$, and $D0.03$ cm^3) are associated with the Y-gradient only and are negative. The 16 scored deviations in Table II are limited to 2 and 3 mm contour spacings. The lowest number (2) is associated with $D5$, both for the 3 mm spacing. As the dose gets higher, closer to the structure surface, the number of errors increases for both $D1$ and $D0.03$ cm^3 (7 each, distributed between 2 and 3 mm contour spacings). With PlanIQ, there were only two deviations above 3%, both with the rotated cylinder with the Y-gradient and 3 mm spacing exhibiting deviations above 3%, where the $D1$ error was −3.9% and $D0.03$ cm^3 error was −4.6%.

3.C. Test 3—Variable contour spacing with variable (matched) dose grid resolution and volume errors across entire curve, per DVH pair

Volume error analysis was performed for 20 dataset permutations: aligned grids for five shapes, two dose gradient orientations, and two resolutions (1 and 3 mm). Results are summarized in Table III. Example DVH curves for four dataset comparisons and for both of the tested systems are shown in Fig. 5. Example graphical histograms comparing volume errors for each system for one dataset (right cylinder, superior/inferior dose gradient, and 1 mm resolution) are plotted in Fig. 6.

4. DISCUSSION

4.A. Treatment of end-capping and resulting deviations

Both PINNACLE (Ref. 10) and PlanIQ (as configured for this work) perform end-capping of the structures consisting of axial contours, extending them by 1/2 axial slice spacing. Both systems include the end-cap volume in the total structure volume calculation.

The treatment of end caps can be easily verified by observing the volume change for the axial cylinder as the contour spacing increases. The additional volume at each end reaches 0.68 cm^3 for 1.5 mm extension of the 24 mm diameter cylinder. However, while PlanIQ (appropriately) includes dose voxels in this extension volume in the DVH calculation, PINNACLE does not. This is evidenced first and foremost by the fact that, for PINNACLE, D_{min} and D_{max} do not change with slice spacing for the Y-gradient (in the direction of the expansion). One hypothesis was that since the axial extension in Tests 2 and 3 is half the size of the dose grid resolution in the Y direction, the voxels outside the drawn contours were somehow ignored in the DVH calculation because of their size in comparison to the extension. The test was repeated with the largest 3 mm slice spacing but the finest available dose grid ($0.4 \times 0.2 \times 0.4$ mm^3), with the same result. Therefore, PINNACLE, while accounting for end-capping in the total volume, systematically ignores dose voxels in the extension regions in the DVH calculation. This is an implementation inconsistency. While there may be good reasons to extend or not to extend the structure drawn on axial slices by 1/2 CT slice thickness, once the decision to expand is made, all dose voxels within the expanded structure should be considered in the DVH. This inconsistency is responsible for the large number of deviations recorded for PINNACLE with the Y dose gradient, particularly in the low dose region (D_{min} and $D99$). As the absolute value of the error rises with increased contour spacing, the deviation above 3%, naturally with the opposite sign, begins to appear near the high end of the DVH, where the denominator is higher (D_{max}, $D1$, $D0.03$ cm^3).

4.B. Dependence on dose grid resolution

The fact that the studied DVH parameters in PINNACLE are independent of the dose grid resolution indicates that the dose is being interpolated for DVH calculations. Otherwise at least D_{min} and D_{max} would vary with dose grid resolution. On the other hand, PlanIQ DVH values—which are also interpolated for all voxels that are between the input dose grid points—are somewhat dependent on the dose grid resolution. This is because PlanIQ will define all voxelated structure volumes to have voxel centers coincident with the original dose grid points, thus assuring that all of those are tallied directly and to avoid skipping over dose peaks or valleys. The input dose grid alignment and resolution, therefore, will impact where the supersampled structure voxels are assigned in 3D space. A practical conclusion can be made that with PINNACLE reducing the CT slice thickness is a more effective way to improve the DVH accuracy for small structures compared to reducing the dose calculation voxel size.

4.C. Observations about D_{min} and D_{max}

A large number of scored PINNACLE deviations, particularly in Test 2 (40), are related to D_{min} and D_{max}. While the

TABLE III. Statistical analysis of the DVH volume (%) error histograms for all datasets at both 1 and 3 mm input data resolution. In all cases, volume (cm^3) differences (numerical–analytical) were calculated for points on the DVH curve sampled at every 0.1 Gy ($N = 301$ for each) then normalized to the structure's total volume (cm^3) to give the error in volume (%).

			DVH volume errors (%), numerical–analytical							
			PINNACLE				PlanIQ			
Dataset	Resolution (mm)	Gradient	Minimum	Maximum	Mean	Standard deviation	Minimum	Maximum	Mean	Standard deviation
Sphere	1	Superior/inferior	−3.0	2.6	0.1	1.2	−0.4	0.4	−0.1	0.1
Sphere	3	Superior/inferior	−9.9	7.7	−0.4	3.5	−1.5	0.0	−0.8	0.6
Sphere	1	Anterior/posterior	−1.8	1.4	0.1	0.8	−0.6	0.4	−0.1	0.1
Sphere	3	Anterior/posterior	−2.8	0.8	−0.4	0.8	−1.9	0.4	−0.8	0.8
Cylinder	1	Superior/inferior	−2.5	2.0	−0.1	1.0	−0.2	0.4	0.1	0.2
Cylinder	3	Superior/inferior	−5.9	5.3	−0.2	2.9	−0.6	0.5	−0.1	0.2
Cylinder	1	Anterior/posterior	−1.9	1.0	−0.2	0.7	−0.2	0.5	0.1	0.3
Cylinder	3	Anterior/posterior	−1.9	0.9	−0.2	0.7	−0.8	0.7	−0.1	0.3
Right cylinder	1	Superior/inferior	−3.6	1.3	−0.9	1.0	−1.0	0.1	−0.4	0.4
Right cylinder	3	Superior/inferior	−9.9	5.0	−2.1	3.3	−4.2	0.0	−2.2	1.0
Right cylinder	1	Anterior/posterior	−3.7	0.0	−1.9	0.9	−1.0	0.8	0.0	0.5
Right cylinder	3	Anterior/posterior	−5.7	0.0	−3.0	1.3	−4.2	0.7	−1.8	1.5
Cone	1	Superior/inferior	−7.0	3.6	0.8	1.8	−0.5	0.5	0.0	0.0
Cone	3	Superior/inferior	−16.7	13.0	0.9	4.4	−0.7	1.4	0.5	0.2
Cone	1	Anterior/posterior	−2.1	1.8	0.5	1.0	−0.3	0.5	0.0	0.0
Cone	3	Anterior/posterior	−2.1	2.1	0.5	1.2	−0.2	0.9	0.5	0.2
Right cone	1	Superior/inferior	−5.8	1.4	−1.7	1.4	0.0	0.6	0.3	0.3
Right cone	3	Superior/inferior	−13.5	8.6	−1.7	3.6	−0.3	1.1	0.3	0.3
Right cone	1	Anterior/posterior	−8.0	0.0	−2.8	1.4	0.0	0.8	0.6	0.3
Right cone	3	Anterior/posterior	−8.3	0.0	−3.0	1.7	−0.3	1.4	0.6	0.3
Average (N = 5)	1	Superior/inferior	−4.4	2.2	−0.3	1.3	−0.4	0.4	0.0	0.2
Average (N = 5)	1	Anterior/posterior	−3.5	0.8	−0.8	1.0	−0.4	0.6	0.1	0.2
Average (N = 5)	3	Superior/inferior	−11.2	7.9	−0.7	3.5	−1.5	0.6	−0.5	0.5
Average (N = 5)	3	Anterior/posterior	−4.1	0.8	−1.2	1.1	−1.5	0.8	−0.3	0.6

variations related to the Y-gradient are explained by the end-capping inconsistency, it is not clear why D_{min} is always 7.5% lower than nominal with the Z-gradient. The DVH algorithm in PlanIQ on the other hand searches for the minimum and maximum dose not only over the voxelated structure grid but also over each original structure surface points defined by the contour coordinates. This postprocessing of the structure surface is why PlanIQ has no errors for D_{min} and D_{max} when compared to analytical results.

It can be argued however that D_{min} and D_{max}, confined to a very small volume, are the least clinically important parameters. Furthermore, D_{min} and D_{max} metrics may show large errors in cases where the volume errors along the same two curves might be very small. After all, D_{min} describes the first dose where a cumulative DVH curve falls below 100%, so if the curves have only a slight gradient at D_{min}, the volume error will by definition be very small. Excluding those parameters from Test 1 would result in PINNACLE scoring 32 (12%) deviations above 3%, compared to 5 (2%) for PlanIQ. Doing the same for Test 2 results in 53 (25%) and 17 (8%) scored deviations, for PINNACLE and PlanIQ, respectively. Thus with the exclusion of D_{min} and D_{max}, the trend remains for the PlanIQ to exhibit fewer deviations when compared to the ideal analytical values.

4.D. Observations about dose gradient direction

Superior/inferior (Y) dose gradients produced more deviations for both systems than the anterior/posterior (Z) dose gradient. This makes sense given that, in addition to the end-capping effect, contour coordinates have high resolution in any axial plane which includes the Z direction, while contour points are discretized to finite values in the Y direction due to slice locations. For dose gradients in the Z direction, any possible errors result purely from the numerical effects: voxel size, voxel grid alignment vs original dose grid locations, dose interpolation, and surface voxel assignment to the inside or outside of the structure.

Excluding the Y-gradient cases and, again, D_{min} and D_{max}, Test 2 scores nine data points outside of the ±3% error interval for PINNACLE and seven for PlanIQ. While numerical performance is similar, all but one PlanIQ deviations are related to the rotated cylinder, while the PINNACLE ones are split between cylinders and cones of different orientation.

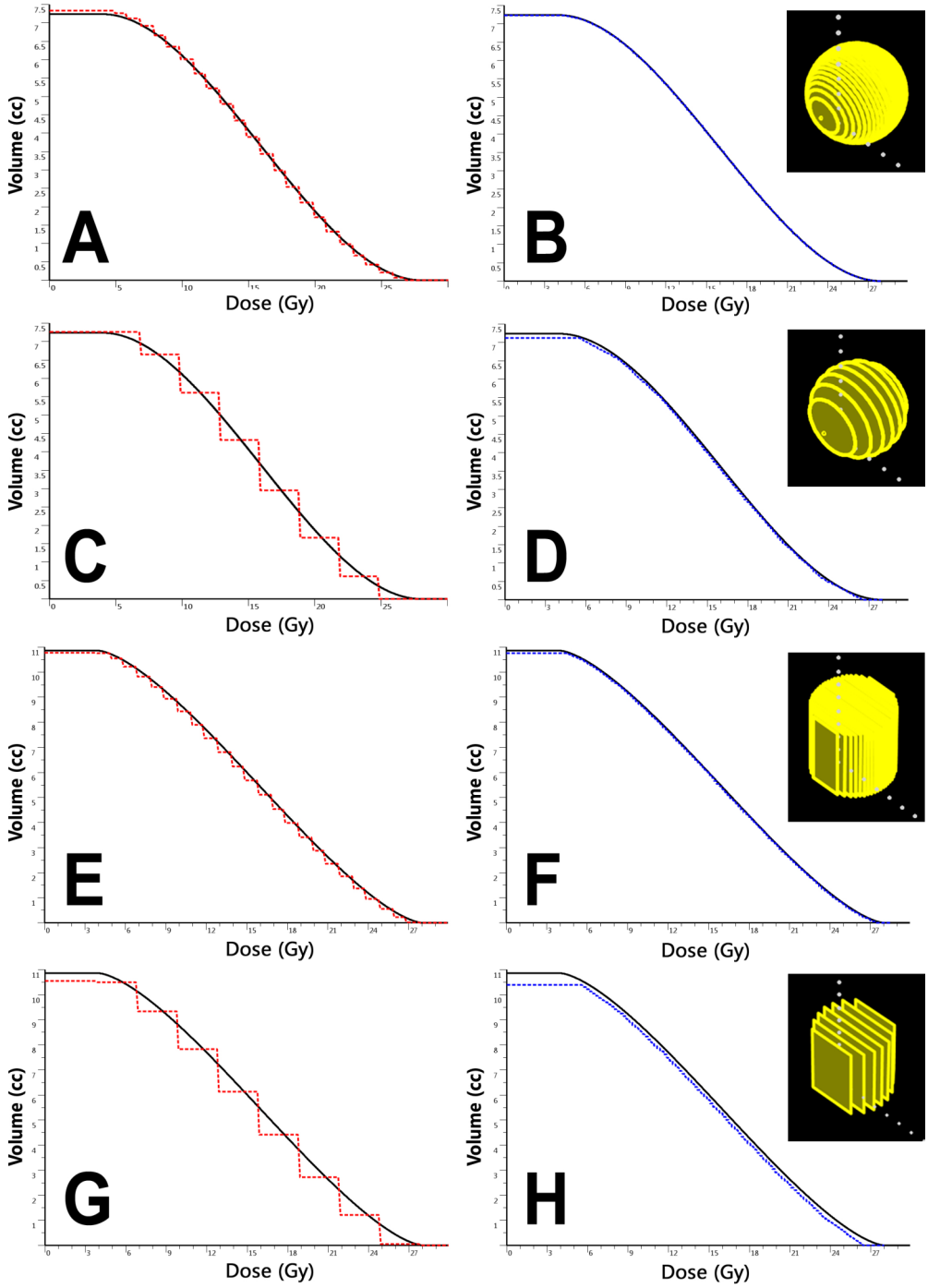

FIG. 5. Example comparisons of analytical (solid lines) vs numerical (dotted lines) for four datasets each with the dose gradient in the superior/inferior dose gradient. In the left hand column [panels (a), (c), (e), and (g)] are the PINNACLE curves vs analytical. In the right hand column [panels (b), (d), (f), and (h)] are the PlanIQ curves vs analytical. Different datasets are each in different rows: row 1 [(a) and (b)]: sphere, 1 mm resolution; row 2 [(c) and (d)]: sphere, 3 mm resolution; row 3 [(e) and (f)]: right cylinder, 1 mm resolution; and row 4 [(g) and (h)]: right cylinder, 3 mm resolution.

Similarly censoring the data for Test 1 results in 16 scored deviations for PINNACLE and five for PlanIQ. Again, all PlanIQ points are associated with the rotated cylinder ($D99$ and $D95$). PINNACLE deviations are split between the cylinders and cones of both orientations, also confined to $D99$ and $D95$ values. Since the effects of the structure discretization are minimized in Test 1, the difference in performance must be attributed to the difference in structure voxelation and dose interpolation between the two systems.

Inspection of Fig. 5 reveals "stair steps" in many curves. These steps occur even though voxelation of interpolated DVH dose is supersampled on the order of 0.1 mm spatial resolution. This is an artifact due to the fact that a linear dose gradient along any one of the orthogonal axes will produce slabs of equal dose voxels (even as thin as 0.1 mm) perpendicular to the dose gradient. If the dose did not change linearly, or at least not along a major axis, there would be no stair steps. It is also worth noting that in Fig. 5 [panels (c) and (g)], the failing

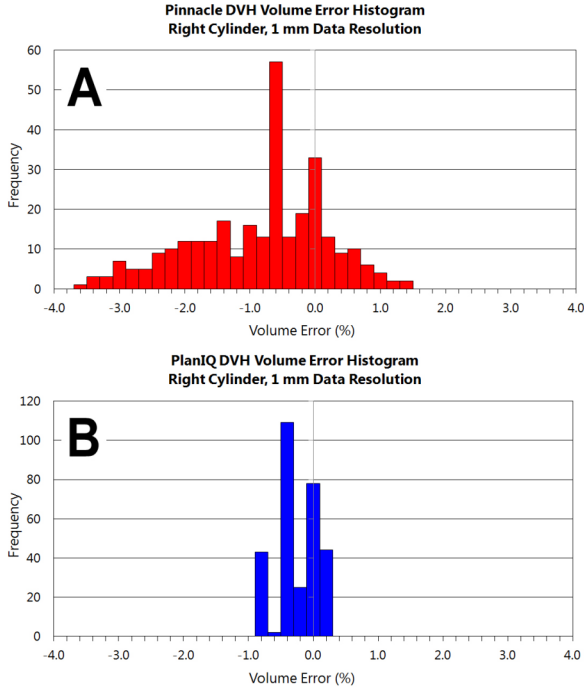

FIG. 6. Example volume error histograms for the right cylinder dataset, superior/inferior dose gradient, and 1 mm input data resolution. Volume errors extracted at 301 points per curve. Panels: (a) PINNACLE errors and (b) PlanIQ errors.

regions are at the high and low dose corners of the DVH, which is not surprising because the dose gradient changes along the undersampled axis (superior/inferior) where both the slice spacing and dose grid resolution (3 mm) are coarse relative to the rapid change in the structure shape at the superior and inferior edges. Since the dose gradient is in this same direction, the limited number of cross-sectional rectangles is not sufficient to model the structure volume. Structure/CT slice spacing is a critical consideration for small structures or for structures whose shape changes rapidly along the Y axis.

4.E. Percent dose difference analysis

As mentioned before, there are no established metrics for DVH accuracy analysis. Since plan evaluation decisions these days are more often not made based on the DVH parameters, it stands to reason that an error in dose (extracted at volume) could be treated with a commonly used 3% dose-difference threshold.[16] However, we note that a more sensitive 2% criterion may be more appropriate for commissioning purposes.[17] Following this logic, we reduced the error threshold to 2%, resulting in 84 (25%) and 11 (3.2%) scored deviations overall for PINNACLE and PlanIQ, respectively. The same procedure for Test 2 yields 106 (42%) and 24 (9%) for PINNACLE and PlanIQ. The trend remains the same as with the 3% threshold.

It is fully understood that the approach taken in this paper is a stress test for any DVH algorithm, particularly given locally normalized percent dose-errors in the low dose region. With $D_{\min} \sim 4$ Gy, the employed dose gradient of 1 Gy/mm equals 25%/mm. Thus, a 3% threshold corresponds to ~0.1 mm spatial difference. This is of course well below the typical CT voxel size, let alone the expected accuracy of structure segmentation.[18]

4.F. Volume error analysis across entire curves

The volume error analysis method done by the CurveCompare software for Test 3 used the extraction of hundreds of points, per DVH curve, to generate statistics and histograms of volume errors. These analyses (Figs. 5 and 6, Table III) reiterate some of the earlier points. In particular, we can conclude that: (1) PINNACLE's lack of sufficient supersampling (voxel resolution) was a limiting factor of its accuracy, and the stair-stepped nature of the DVHs for these small structures translates to a higher range and variation of volume errors across any given curve compared to the PlanIQ's smoother curves and tighter, less variable error histograms; (2) lack of data resolution (dose grid and CT spacing) hinders DVH accuracy and is exacerbated for dose gradients in the superior/inferior direction; and (3) structures that change shape rapidly in the superior/inferior direction are more difficult to voxelize accurately vs shape changes in axial plans (L/R, A/P) which are not limited in resolution by discrete axial plane locations.

It is worth noting that rigorous DVH testing (especially of a numerical calculation vs an absolute analytical curve) should analyze dose and volume errors, as in clinical practice, we extract both dose-at-volume and volume-at-dose metrics. One may consider a method akin to the "gamma"[19] metric applied to the DVH,[8] however, when validating the accuracy of a system, it is best to use the most sensitive methods available in order to diagnose systematic errors no matter how small, with the hopes to improve the system and drive out such errors. A gamma analysis type combination metric will always decrease sensitivity, whereas studying dose and volume errors separately will give the full picture without potentially hiding points where the dose (or volume) error might be very high even though the curves are in close proximity in that region of the curve but along a very steep gradient. For instance, $D99$ target volume errors can be quite large even if curves look very similar, and this is important considering that often $D99$ or $D95$ volumes are taken alone when evaluating target coverage.

4.G. Applications and discussion of analysis metrics and acceptance criteria

The aim of this work is to provide benchmark analytical datasets and introduce several methods that can be used to validate accuracy of the DVH curves and statistics produced by medical software systems (e.g., TPS). We elected to extract and analyze both dose error and volume error statistics and believe that one cannot choose solely one or the other. Both are important by virtue of the fact that critical DVH points used in plan evaluation can be either type, such as $D0.03\ \text{cm}^3$ (for study of spinal cord clinically relevant maximum dose) or volume (%) of a target receiving the prescription dose. Since plan evaluation is often based on discrete, extracted points,

then dose error and volume error analyses are both critical during commissioning. Tests 2 and 3 use the same dataset permutations but different analysis methods and metrics, with Test 2 being a more general analysis of a system's accuracy over ranges of input data quality, while Test 3 provides more thorough, or diagnostic, analysis per DVH pair, per dataset.

With these well-controlled standard datasets and specialize software now in hand and generally available, the natural question to follow is "What is acceptable in terms of DVH agreement?" The principle aim of this paper is not to be prescriptive in answering this question (though this might be an interesting topic for a task group) but rather to provide useful data and methods with which to study the agreement metrics in the first place. However, we can offer some reflections based on our experience. We summarize these below.

- As stated earlier, we suggest that dose errors and volume errors be considered separately and equally weighted. The reason is because clinicians rely on both dose-at-volume and volume-at-dose metrics in plan evaluation. "Gammalike" curve closeness factors combining dose and volume errors[8] may hide high, local errors at critical points.
- During acceptance testing/commissioning, the more sensitive the metrics and more stringent the tolerances, the more can be learned.
- Pay attention to any errors (dose and volume) that exceed 2%. In addition, pay attention to the location of those regions. Are they along gradients? Are they due to stair-stepping/resolution issues? Are they reproducible across multiple volume shapes and orientations?
- It's reasonable to suggest that the strategic, discrete set of points extracted in Test 2 (dose-at-volume) could be replaced with the general Test 3 method of sampling dose errors continuously along the entire curve. CurveCompare software allows analysis of errors on either axis. However, as a hybrid, one might want to filter the dose or volume range analyzed to avoid large regions where even inaccurate DVH curves should overlap.
- There is merit in the volume- (or dose-) error histograms, but visual inspection of curves with a keen eye is also quite valuable.
- Noteworthy errors (most importantly systematic ones) can be reported to the DVH software's vendor so that they can run the same tests with the same data to diagnose root causes. In turn, the vendor can fix the issue(s) and release new software to improve accuracy to the entire user base. Ideally, the vendors would adopt these test methods as part of product validation, prior to commercial release, and publish results with their product documentation so that users can see performance benchmarks.

4.H. Continual improvement and the driving out of DVH errors

It is important to clarify that errors in DVH calculations are mostly avoidable. It is sheer mathematics, without the inherent uncertainties we are used to considering with, for example, absolute calibration or dose calculation. One of the problems is that in some commercial systems DVH calculations were coded many years ago and not revisited. The limitations on resolution imposed by the available computing power then are likely irrelevant today. The DVH calculational accuracy is potentially limited only by the coarseness of the contour resolution, particularly in the superior/inferior direction.

It is our hope that using our data and methods (plus any additional datasets that complement ours), other groups will analyze the accuracies of their respective systems and intercompare. With enough data accrued, it will become feasible to study results and prescribe standard metrics and suggest stringent, but reasonable, acceptance criteria. Potentially, a DVH software system that shows the highest accuracy against analytical datasets could be used to produce reference numerical DVH curves for clinically relevant structures and dose distributions, i.e., to act as surrogate when analytical curves are not possible.

5. CONCLUSIONS

Analytical formulas were derived to calculate cumulative DVHs of the simple geometrical 3D objects in the presence of the dose distributions with 1D linear gradient. Corresponding numerical DICOM RT structure and dose objects were developed programmatically as input to the DVH calculation algorithms. Custom analysis software was developed to study the important metrics of DVH accuracy. Together, these elements constitute a method which enables the measurement of the performance of any DVH calculation system against ideal analytical values.

The methods, designed as algorithm stress tests, were used to evaluate two software packages—PINNACLE and PlanIQ. Within the parameters of the tests, PlanIQ exhibited overall fewer deviations from the expected analytical values than PINNACLE. A large number of deviations in PINNACLE are driven by an inconsistency in implementation of end-capping. While the total structure volume is increased by the superior/inferior (Y) extension by 1/2 CT slice thickness, the dose voxels falling within that extension are apparently not included in the DVH calculation. Excluding data points prone to the end-capping effect (Y-gradient) as well as $D_{\min}$ and $D_{\max}$, PlanIQ still exhibits fewer deviations from the ideal values, particularly with the test where the contour axial spacing was kept at a minimum (0.2 mm) while the dose grid resolution varied.

In general, the method proved useful in measuring performance and identifying failure modes of DVH calculation and can be universally applied to the testing of any software system that calculates DVHs.

ACKNOWLEDGMENTS

This work was supported in part by a grant from Sun Nuclear Corporation (SNC). B.N. is a consultant to SNC; however, his contribution to this work was outside that consultancy. CurveCompare software is noncommercial.

APPENDIX: DERIVATION OF DVH FOR ROTATED CONE

Here, the analytical formula for the cumulative DVH of a rotated cone in the presence of a linear 1D dose gradient is derived. While DVH derivation for all structures follows the same logic, this one is mathematically the most challenging. The object is a cone centered on the coordinate system origin [Fig. 2(d)], having a height of H and a base radius of R. The height is aligned with the Z-axis and extends from $z=-H/2$ to $H/2$. The equation defining the structure contour on an axial slice is that of a hyperbolic cone section,

$$\frac{R^2(z+H/2)^2}{H^2y^2}-\frac{x^2}{y^2}=1. \tag{A1}$$

The variables z and x can be parameterized as

$$\begin{aligned} z&=\frac{Hy}{R}\cosh(\omega)-\frac{H}{2}\\ x&=y\sinh(\omega). \end{aligned} \tag{A2}$$

An element of area dA on the axial slice can be written as $dA=xdz$. To obtain axial cross-sectional area A, it needs to be integrated over z from $(Hy/R–H/2)$ to $H/2$. However, it is advantageous to integrate instead over ω, whose limits are from 0 to $\operatorname{arcosh}(R/y)$,

$$\begin{aligned} A&=2\int_0^{\operatorname{arcosh}(R/y)}\frac{H}{R}y^2\sinh^2(\omega)d\omega\\ &=HR\left(\sqrt{1-\frac{y^2}{R^2}}-\frac{y^2}{R^2}\operatorname{arsech}\left(\frac{|y|}{R}\right)\right). \end{aligned} \tag{A3}$$

The factor of 2 in the integral for the area is required to account for both halves of the hyperbolic curve. The running volume as a function of y, $V(y)$, can then be determined as an integral of the area A over y,

$$\begin{aligned} V(y)&=\int_y^R HR\left(\sqrt{1-\frac{y^2}{R^2}}-\frac{y^2}{R^2}\operatorname{arsech}\left(\frac{|y|}{R}\right)\right)dy\\ &=HR^2\left(\frac{\pi}{6}-\frac{2}{3}\frac{y}{R}\sqrt{1-\frac{y^2}{R^2}}-\frac{1}{3}\arcsin\left(\frac{y}{R}\right)\right.\\ &\quad\left.+\frac{1}{3}\frac{y^3}{R^3}\operatorname{arsech}\left(\frac{|y|}{R}\right)\right). \end{aligned} \tag{A4}$$

At the same time, dose D varies linearly in the y-direction, from maximum $D_{\max}$ at $y=R$ to minimum $D_{\min}$ at $y=-R$ and can be expressed in terms of $\Delta D=D_{\max}-D_{\min}$ and $\Sigma D=D_{\max}+D_{\min}$,

$$D(y)=\frac{\Delta D}{2R}y+\frac{\Sigma D}{2}. \tag{A5}$$

Solving for dose,

$$y(D)=\frac{R}{\Delta D}(2D-\Sigma D). \tag{A6}$$

Substituting Eq. (A6) into Eq. (A4), the formula for normalized cumulative DVH is finally obtained as

$$\begin{aligned} V(D)&=\frac{1}{2}-\frac{2}{\pi}\frac{2D-\Sigma D}{\Delta D}\sqrt{1-\frac{(2D-\Sigma D)^2}{\Delta D^2}}\\ &\quad-\frac{1}{\pi}\arcsin\left(\frac{2D-\Sigma D}{\Delta D}\right)\\ &\quad+\frac{1}{\pi}\frac{(2D-\Sigma D)^3}{\Delta D^3}\operatorname{arsech}\left(\frac{|2D-\Sigma D|}{\Delta D}\right). \end{aligned} \tag{A7}$$

Volume at a given dose is thus expressed analytically. To determine the minimal dose covering a given volume ($D95$, etc.), this equation can be easily numerically solved for D with necessary precision.

[a]Author to whom correspondence should be addressed. Electronic mail: vladimir.feygelman@moffitt.org

[1]M. M. Austin-Seymour, G. T. Y. Chen, J. R. Castro, W. M. Saunders, S. Pitluck, K. H. Woodruff, and M. Kessler, "Dose volume histogram analysis of liver radiation tolerance," Int. J. Radiat. Oncol., Biol., Phys. **12**, 31–35 (1986).
[2]R. E. Drzymala, R. Mohan, L. Brewster, J. Chu, M. Goitein, W. Harms, and M. Urie, "Dose-volume histograms," Int. J. Radiat. Oncol., Biol., Phys. **21**, 71–78 (1991).
[3]G. T. Y. Chen, "Dose volume histograms in treatment planning," Int. J. Radiat. Oncol., Biol., Phys. **14**, 1319–1320 (1988).
[4]B. Fraass, K. Doppke, M. Hunt, G. Kutcher, G. Starkschall, R. Stern, and J. Van Dyke, "American association of physicists in medicine radiation therapy committee task group 53: Quality assurance for clinical radiotherapy treatment planning," Med. Phys. **25**, 1773–1829 (1998).
[5]E. Panitsa, J. C. Rosenwald, and C. Kappas, "Quality control of dose volume histogram computation characteristics of 3D treatment planning systems," Phys. Med. Biol. **43**, 2807–2816 (1998).
[6]H. Chung, H. Jin, J. Palta, T. S. Suh, and S. Kim, "Dose variations with varying calculation grid size in head and neck IMRT," Phys. Med. Biol. **51**, 4841–4856 (2006).
[7]J. F. Corbett, J. Jezioranski, J. Crook, and I. Yeung, "The effect of voxel size on the accuracy of dose-volume histograms of prostate 125I seed implants," Med. Phys. **29**, 1003–1006 (2002).
[8]M. A. Ebert, A. Haworth, R. Kearvell, B. Hooton, B. Hug, N. A. Spry, S. A. Bydder, and D. J. Joseph, "Comparison of DVH data from multiple radiotherapy treatment planning systems," Phys. Med. Biol. **55**, N337–N346 (2010).
[9]W. Straube, J. Matthews, W. Bosch, and J. Purdy, "SU - FF - T - 310: DVH analysis: Consequences for quality assurance of multi - institutional clnical trials," Med. Phys. **32**, 2021–2022 (2005).
[10]T. Ackerly, J. Andrews, D. Ball, M. Guerrieri, B. Healy, and I. Williams, "Discrepancies in volume calculations between different radiotherapy treatment planning systems," Australas. Phys. Eng. Sci. Med. **26**, 91–93 (2003).
[11]M. A. Ebert, A. Haworth, R. Kearvell, B. Hooton, R. Coleman, N. Spry, S. Bydder, and D. Joseph, "Detailed review and analysis of complex radiotherapy clinical trial planning data: Evaluation and initial experience with the SWAN software system," Radiother. Oncol. **86**, 200–210 (2008).
[12]W. L. Straube, W. R. Bosch, J. W. Matthews, S. M. Goddu, W. C. Bennett, J. M. Michalski, and J. A. Purdy, "An electronic phantom for testing and quality assurance of dose volume histogram calculations used in data analysis for multi institutional clinical trials," Int. J. Radiat. Oncol., Biol., Phys. **75**, S620–S621 (2009).
[13]E. Panitsa, J. C. Rosenwald, and C. Kappas, "Developing a dose-volume histogram computation program for brachytherapy," Phys. Med. Biol. **43**, 2109–2121 (1998).
[14]J. L. Bedford, P. J. Childs, V. N. Hansen, M. A. Mosleh-Shirazi, F. Verhaegen, and A. P. Warrington, "Commissioning and quality assurance of the PINNACLE 3 radiotherapy treatment planning system for external beam photons," Br. J. Radiol. **76**, 163–176 (2003).
[15]See supplementary material at http://dx.doi.org/10.1118/1.4923175 for the complete data spreadsheet, DICOM RT objects, and CurveCompare software.
[16]G. A. Ezzell, J. W. Burmeister, N. Dogan, T. J. LoSasso, J. G. Mechalakos, D. Mihailidis, A. Molineu, J. R. Palta, C. R. Ramsey, B. J. Salter, J. Shi, P. Xia, N. J. Yue, and Y. Xiao, "IMRT commissioning: Multiple institution planning and dosimetry comparisons, a report from AAPM task group 119," Med. Phys. **36**, 5359–5373 (2009).

[17]B. E. Nelms, M. F. Chan, G. V. Jarry, M. Lemire, J. Lowden, C. Hampton, and V. Feygelman, "Evaluating IMRT and VMAT dose accuracy: Practical examples of failure to detect systematic errors when applying a commonly used metric and action levels," Med. Phys. **40**, 111722 (15pp.) (2013).
[18]X. A. Li, A. Tai, D. W. Arthur, T. A. Buchholz, S. Macdonald, L. B. Marks, J. M. Moran, L. J. Pierce, R. Rabinovitch, A. Taghian, F. Vicini, W. Woodward, and J. R. White, "Variability of target and normal structure delineation for breast cancer radiotherapy: An RTOG multi-institutional and multiobserver study," Int. J. Radiat. Oncol., Biol., Phys. **73**, 944–951 (2009).
[19]D. A. Low, W. B. Harms, S. Mutic, and J. A. Purdy, "A technique for the quantitative evaluation of dose distributions," Med. Phys. **25**, 656–661 (1998).
[20]In PlanIQ supersampling is a preference that is turned "on" by default. It can be toggled "off" by the user but we sought to validate the software as we use it, and we would never disable supersampling.
[21]CurveCompare is noncommercial research software and is available to the research community along with the analytical data and DICOM structure sets created for this work (see Ref. 15).