Analysis and evaluation of planned and delivered dose distributions: practical concerns with γ- and χ- Evaluations

2013 J. Phys.: Conf. Ser. 444 012016

(http://iopscience.iop.org/1742-6596/444/1/012016)

Analysis and evaluation of planned and delivered dose distributions: practical concerns with γ- and χ- Evaluations

LJ Schreiner[1,2], O Holmes[2] and G Salomons[1,2]
[1]Cancer Centre of Southeastern Ontario at Kingston General Hospital, 25 King Street West, Kingston, ON, Canada, K7L5P9
[2]Department of Physics, Queen's University, Kingston, Ontario, Canada, K7L3N6

E-Mail: john.schreiner@krcc.on.ca

Abstract. One component of clinical treatment validation, for example in the commissioning of new radiotherapy techniques or in patient specific quality assurance, is the evaluation and verification of planned and delivered dose distributions. Gamma and related tests (such as the chi evaluation) have become standard clinical tools for such work. Both functions provide quantitative comparisons between dose distributions, combining dose difference and distance to agreement criteria. However, there are some practical considerations in their utilization that can compromise the integrity of the tests, and these are occasionally overlooked especially when the tests are too readily adopted from commercial software. In this paper we review the evaluation tools and describe some practical concerns. The intent is to provide users with some guidance so that their use of these evaluations will provide valid rapid analysis and visualization of the agreement between planned and delivered dose distributions.

1. Introduction

Improved treatment planning systems, dose delivery equipment (typically, linear accelerators) and therapeutic techniques have advanced modern radiation therapy by enabling clinicians to achieve more conformal high dose delivery to target volumes while sparing adjacent normal tissues. However, as discussed throughout these proceedings, these improvements have also increased significantly the requirement for dose delivery validation especially in the commissioning of treatment units and new treatment techniques and, in some settings, in the patient specific validation of dose delivery.

The measurement of the integrity of dose delivery during the commissioning of a new unit, or a new treatment technique, may involve the measurement of dose distributions in phantoms for test cases planned under well-defined conditions [1,2]. Such measurements can ensure correct performance and establish benchmark data for future quality assurance of the particular treatment protocol. The measurements often involve film or two or three dimensional (2D and 3D) diode or ion chamber array measurements on regular and anthropomorphic phantoms. These devices may also be used to test the delivered distributions for specific patients by exposing the devices in phantom to the same multileaf collimator leaf sequences, trajectories and MUs as planned for the treatment [3]. Patient treatment delivery validation may also be performed using dose reconstruction from exit beam measurements on Electronic Portal Imaging Devices (EPIDS) [4]. And finally whole treatment protocols may be regularly monitored in end to end test procedures using 2D and 3D dose delivery validation [5,6]. Gel and radiochromic 3D dosimetry has played a role in all these settings in select clinics.

2. Evolution of dose evaluation: introduction of dose and distance metrics

In all the situations above, the comparison of calculated dose distributions to physical measurement is a major task, since both the reference and the measured distributions usually consist of large 2D or 3D data sets [2,7,8,9]. Furthermore, the very nature of the comparison of dose distributions requires that the evaluation probe both dose and spatial domains [8]. An approach of past decades to accommodate dosimetric and spatial information by overlaying 2D dose contours in various planes for comparison was sufficient if the contours agreed, but the delivery was not easily evaluated if agreement failed. A better alternative for the comparison has been to calculate a dose difference map: a spatial representation of the numerical difference in the doses in the two distributions. There is a shortcoming with this technique in that the approach is inherently oversensitive in regions of high dose gradient where small spatial errors in either data set can lead to large dose differences between the measured and planned distributions [7,8,9]. Separate methods, such as the use of the distance to agreement (DTA) which reports the distance between the calculated point and the nearest point in a measured dose distribution with the same dose value [8], are needed to obtain useful comparisons in regions of high dose gradient. However, while DTA maps improve the dose comparisons in regions of high dose gradient, they tend to be overly sensitive in regions of uniform dose (low dose gradients) such as in the target volume.

A solution to these problems has been developed over the last 15 years through the introduction of various metrics combining agreement criteria for dose and distance in the distributions being compared [7,8,9]. Harms *et al*. [8] combined the complementary nature of the DTA and dose difference distributions in a software tool called the composite evaluation. In this evaluation, the dose difference and DTA distributions were independently evaluated as either meeting or exceeding predefined tolerances. Harms [8] used a dose difference criterion of 3 % (i.e., 3 % of the maximum dose) and a 3mm limit for the DTA, this was slightly different than the 3%/4mm tolerances suggested by Van Dyk for the evaluation of treatment planning systems [11]. The 3%/3mm criteria has been subsequently adopted in many other studies [9,12-18]. In Harms composite evaluation, the tolerance criteria were used to create two binary pass-or-fail distributions (corresponding to the DTA and dose difference) which were multiplied point by point to give a single binary distribution. The combined binary distribution contained a binary array (0's or 1's) indicating agreement or failure, respectively, within 3%/3mm. The final output of the evaluation showed regions which failed to meet both the dose and DTA criteria and, therefore, it did not suffer from increased sensitivity in regions of low or high dose gradient. But as it was only a binary map, the evaluation visualizing the outcome was not easily interpreted [8,9]. Also the approach heightened the impression of disagreement in regions of high dose gradient and was not representative of the significance of agreement in these regions.

3. Gamma dose evaluation

Low [7,9,12] formalized the approach of evaluating both dose and spatial characteristics of the two distributions being compared by the development of the gamma (γ) dose distribution method. A γ comparison is performed between two dose maps: one distribution is the 'reference plan' (typically from the treatment planning system, but see discussion below) and the other is the 'evaluated distribution', usually from a two or three dimensional dose measuring system. The reference distribution is treated as the true distribution, while the evaluated image is analyzed for its agreement with the reference as follows: every point in the reference image has a corresponding γ value; a measure of agreement at that location. Each possible point in the reference distribution can be coupled with any point in the evaluated image. For all pairs there exists a Γ, defined by the vector difference between the points.

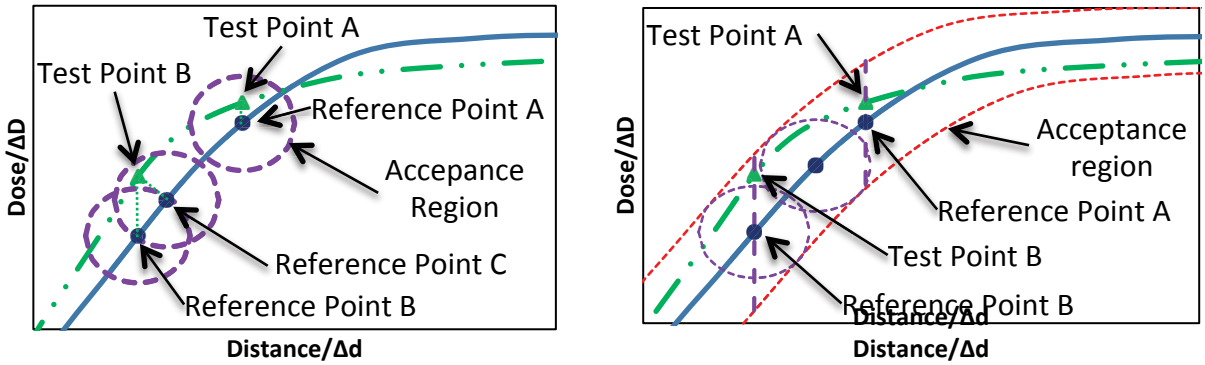

Figure 1: a) An illustration of a 1D γ test. For each reference point (blue dot) there is a circular acceptance region (purple dashed line). The γ test for each reference point involves a search to find a point on the evaluation curve (a test point) that falls within the reference point's acceptance circle. The passing test point is not necessarily at the same physical resolution as the reference point (as shown there are only two test points on the evaluation curve, marked by green triangles). b) An illustration of a 1D χ test. For each reference point (blue dot) an extended dose limit criteria is calculated (purple dashed line) based on the dose gradient and distance tolerances at that point. The extended dose limits essentially indicate the upper and lower limits of the acceptance region surrounding the reference curve. The χ test for each reference point involves an interpolation of the test curve points to the positions of the reference points and a comparison of the interpolated test dose values to the extended dose limit criteria.

Tolerance criteria Δd_M and ΔD_M (e.g., 3mm and 3%) are used to normalize Γ along the distance and dose vector dimensions, correspondingly. The gamma index value, γ, is the smallest Γ that can be found considering the entire evaluated distribution. In 2D evaluation space $\Gamma^2 = 1$ describes a circle whose extent is defined by the tolerance criteria (see figure 1a), whereas in a 3D space the criteria define an ellipsoid. When γ≤1, the distributions agree within the stipulated tolerances. Conversely, when γ>1, no point in the evaluated dose map can be found within the circle/ellipsoid and the dose distributions disagree at that location. As with the composite test [14], the γ function identifies if the the evaluated distribution passes (γ ≤1) or fails (γ>1) the comparison. Also, if the evaluated distribution fails, the value of γ indicates roughly the degree of failure relative to the dose and distance to criteria set in the evaluation [7]. For example, a γ = 1.5 fails by 50% which corresponds to a failure of 1.5% or 1.5mm for a 3%/3mm dose/DTA criteria; or 1% or 1mm for 2%/2mm criteria. If the vector nature of the gamma is evaluated [7,15,19] the test can give indications of whether the failure is primarily due to dose or distance to agreement failure.

4. Chi dose evaluation

Since it was introduced, the γ-tool has been refined, modified and evaluated by several authors [12-20]. One alternative to the γ function suggested was the chi (χ) evaluation proposed by Bakai *et. al* [14], which also provides a metric combining both dose and DTA criteria to evaluate the level of agreement between evaluation and reference dose distributions. However, in the χ approach the comparison between the reference and evaluated distributions is carried out differently than with the γ test. Instead of searching the test space for evaluation points that are closest to a given reference point, the χ test compares the reference dose to the test dose at the same point in space, but it scales the dose limit criteria according to dose gradient of the reference distribution.

The difference between γ and χ evaluations is illustrated in figure 1. Discrete points on the reference curve are systematically compared (see figure 1a) to discrete points on the evaluation curve. In the figure test point A on the evaluation curve is within the acceptance region for reference point A. The test point B is not within the acceptance region for reference point B, but then it is within the acceptance region for reference point C. This illustrates that the γ evaluation must search multiple test points to find a minimum Γ. When calculating gamma over a volume, this can become

computationally expensive. The χ test takes a different approach. Instead of identifying an acceptance region about individual reference points, and determining if evaluation points fall within one of these regions, the χ test extends the dose tolerance based on the dose gradient and distance limit at that point. This is illustrated in figure 1b, where the χ dose limits, shown as the vertical dashed lines, have been extended beyond the dose-only tolerance (the circle) to reflect the limits of the "acceptance tube" surrounding the reference curve. In this way, the test points are each compared to the extended dose limits for that point with no need to search.

Beyond the fact that there is no need to search the test space, there are several other advantages to the χ test. First, the χ test returns a positive or negative value indicating whether a failure is an

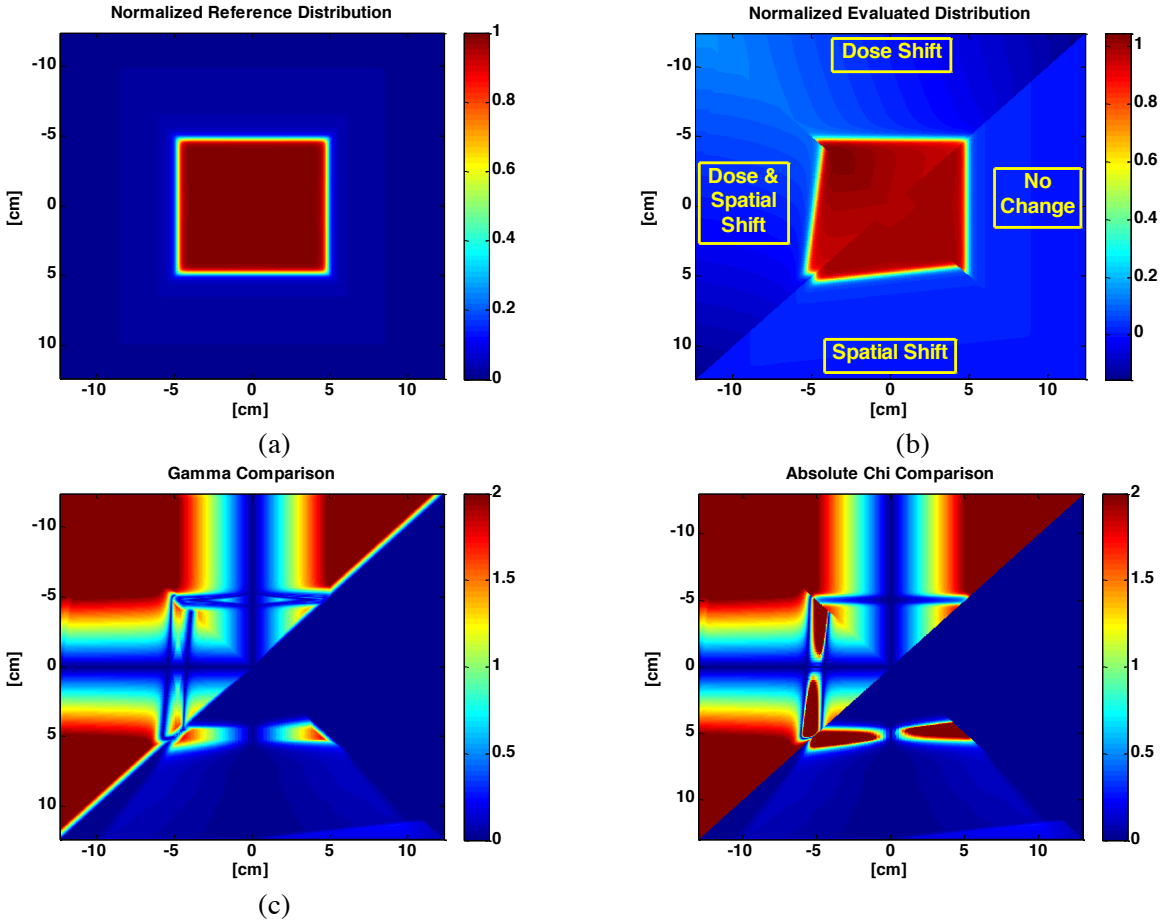

Figure 2: An illustration of the output from γ and χ comparisons between two test distributions that were first used by Low and Dempsey [13] to illustrated properties of γ evaluations. (a) The reference distribution is a normalized 10 cm by 10 cm square field originally from a 6MV irradiation. (b) The evaluated dose distribution is divided into quadrants along the diagonals, and each quadrant is perturbed from the original reference distribution in different ways as follows: the region at noon has the dose shifted by 0.12*x* % (*x* measured in mm), at three o'clock the distribution is not modified and so the evaluated and reference distributions correspond exactly at all points, in the quadrant at six o'clock , the points in the distribution have been shifted spatially by 0.12*x* mm (*x* measured in mm), and at nine o'clock, both shifts have been implemented (with the effect altered to depend on the *y* dimension). (c) Shows the γ comparison map calculated with the original γ-tool (reproduced with permission). (d) Shows distribution of the absolute value of the calculated χ [14] with an in-house χ-tool. (Note that the colour bar on the right presents the scale for the results of the dose comparisons, herein and in other figures below).

overdose or an under-dose (although this is not easily visualized in colour wash and so representations of the result of the χ evaluation, such as figure 2d, usually show absolute values.). In some situations, this allows one to more easily evaluate if a failure is clinically relevant (e.g., if it is an acceptable

under-dosing in the region of an organ of risk). Another advantage with the χ evaluation is that one can assign different dose and distance limits in different regions of the dose distributions being compared and can also allow the dose criteria in the positive and negative directions to differ. This would enable one to apply tighter dose and/or distance criteria to the region around an organ of particular concern and looser criteria elsewhere.

The χ evaluation does have some limitations. Firstly, it requires that the points on the evaluation distribution are spatially matched to the reference points. This means that the data for the evaluation dose distribution may need to be interpolated, with the potential for some distortion. Secondly, the extended dose limits (i.e., establishing the acceptance regime outlined by the red curves in figure 1b) in regions where the dose gradient changes rapidly (where the second derivative is large). The result is that χ values near the shoulder portion of an evaluated curve may not be calculated accurately. In spite of these limitations, the χ test is a strong alterative to the traditional γ evaluation and it has recently been gaining broader clinical use.

5. Practical Considerations I: discretization and spatial resolution

While Low *et al*. presented a powerful theoretical concept for the evaluations based on continuous functions [9], real clinical comparisons are typically made between discretized representations of dose distributions often with the reference and evaluation dose data sampled at different spatial resolutions. The importance of resolution was first analysed by Depuydt *et al*. [12] in their clinical assessment of the γ evaluation. They were particularly concerned with overestimations of γ values caused by large grid spacing in the discrete dose distributions, particularly in regions of high dose gradient. To avoid overestimating they effectually reduced the continuous value of γ to a binary test equivalent to the composite evaluation [8]. Several other studies have been directed, in part, towards resolving this issue [14,17,18]. Low and Dempsey [13] noted that by re-sampling the distributions to 1x1mm2 grids, the error in γ associated with pixelization artifacts was reduced to less than 0.2, even in regions of high dose gradient. Others have minimized discretization artifacts by interpolating additional data points in the evaluated grid [17].

The effect of resolution on the resulting γ distributions is illustrated in figure 3, showing how the γ comparison maps and the observed pass rate (a common parameter used to roughly summarize the complex γ dose distribution comparison by reporting the percentage of points in the dose distribution that pass the specified γ criteria Δd_M and ΔD_M) change with spatial sampling (the bottom rows corresponding to different resolutions). When the film data (fixed at a resolution of 0.24 mm) are taken as the reference doses, and the Eclipse plan is set as the evaluation distribution, increasing the resolution of the evaluated distribution (from 2.5 to 0.24 mm) changes the pass rate from 80.9% to 91.3%. This result can be explained by the behaviour of the γ search. When the evaluated distribution has a coarse pixel size compared to the reference distribution, many reference pixels fall a significant distance from the nearest evaluated pixel. Thus, the γ value for many reference pixels reflect significant spatial misalignment purely as an artefact of the coarse evaluated resolution which is seen as the grid-lines in the low dose region of figure 3 (c). When the resolution of the evaluated distribution is increased to match that of the reference distribution, this spatial artefact is eliminated since each reference point has a directly corresponding pixel in the evaluated distribution. Increasing the evaluated resolution also provides each reference point with a greater range of dose values for comparison, further increasing the likelihood of finding a set of pixels which pass the gamma test. Note, however, that the change of resolution in the Eclipse plan does not significantly change the pass rate if the Eclipse plan is the reference distribution. This is discussed further below.

The requirement for interpolation adds a significant increased computational burden to γ evaluations [17] and the computation time grows as a third power of increasing grid resolution [18]. The computational speed of the algorithm can be enhanced by pre-calculating interpolation factors [17]. Another approach to reducing the evaluation time is to restrict searches around each reference point [15, 17, 18, 19]. For example, one can limit searches *a priori* so that points that "do not have a

chance" of yielding the smallest value of Γ are eliminated: as soon as $\Delta d/\Delta d_M$ becomes larger than the smallest Γ found so far, the search is terminated. Others [18] have developed geometric approaches to determining the smallest Γ.

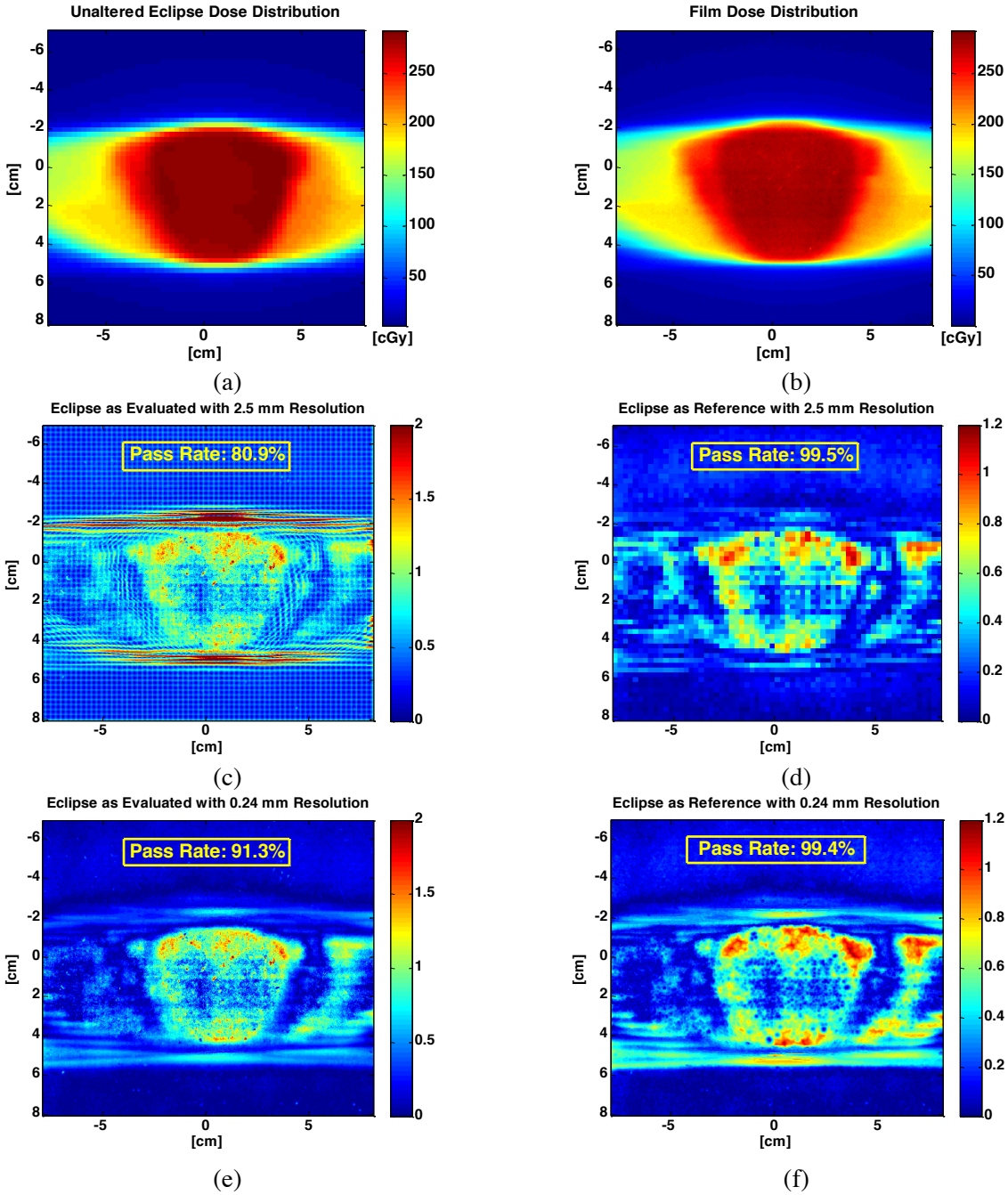

Figure 3: An illustration of the importance of two properties of the γ evaluation: i) the spatial resolution of the data sampling in the evaluated and reference dose maps, and ii) the status of the two distributions as reference or evaluated distributions. The top row shows (a) the calculated distribution for a prostate treatment plan from the Eclipse planning system and (b) a measurement of this dose distribution using radiochromic film at a 0.24 mm resolution. The middle row shows the γ comparison map from film data at the original 0.24 mm resolution and using the Eclipse data at the original 2.5 mm resolution as (c) the evaluated distribution and (d) the reference distribution. The bottom row shows the γ distribution from film data at the original 0.24 mm resolution and the Eclipse data interpolated to a 0.24 mm resolution as (e) the evaluated distribution and (f) the reference distribution.

6. Practical Considerations III: the role of reference and evaluation distributions

In the previous discussion of figure 3, it was noted that the results of the comparison of two dose distributions is affected by the designation of the distributions as reference or evaluated data sets. Neither γ evaluations, nor DTA map, are symmetric with respect to the distributions being compared. Before this is discussed in detail, it is important to note an important change of notation introduced by Low and Dempsey [13] to recognize that in clinical practice the distributions can both be calculated (e.g., output from two different planning systems), or from two separate measurements (e.g., on alternative dosimetric systems being evaluated for performance). For example, it may be necessary to perform comparisons between Monte Carlo computations and calculations made by commercial planning software, or between film measurements. So to avoid confusion, the terms 'reference' and 'evaluated' distributions were adopted by Low to replace 'measured' and 'calculated' distributions, respectively. This notation has been become widely adopted [15, 17, 18]. The convention is different than that used historically, for example, during the development of the DTA and dose difference tools in the evaluation of electron beam dose calculation software [10]. In the current usage it is usually understood that the dose comparison is used to measure the extent to which an evaluated distribution agrees with a reference distribution, which is treated as the *true* distribution (though, strictly speaking, obtaining a true distribution may not be possible).

That the designation of which distribution is set as the reference can affect the results of the γ comparison was illustrated previously in figure 3, where the role of the Eclipse dose distribution is switched in the bottom columns. This is also true for the χ comparisons (see figure 4). As noted above, part of the change in the γ comparisons can be related to differences in resolution. Noise in the data sets also contributes to the differences [7, 13]. In the examples shown in figures 3 and 4, the film dose distribution data are noisier and the Eclipse data smoother.

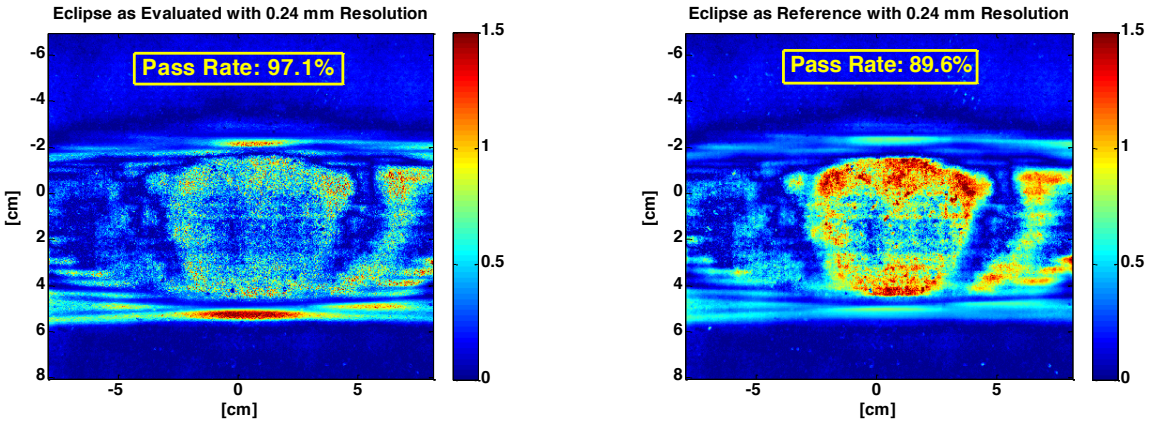

Figure 4: An illustration of the importance of the status of the two distributions as reference or evaluated distributions for the χ evaluation comparing the Eclipse and film dose distributions shown in figure 3a and 3b, respectively. (a) the χ distribution map with the Eclipse data is the evaluated distribution. (b) the χ distribution with the Eclipse data is the reference distribution.

In figure 3 the side-by-side comparison maps (a)-(b) and (c)-(d) show substantial increases in pass rate when the noisy film data forms the evaluated dose distribution. This result stems from the search algorithm of the γ function which compares points in the reference distribution to all of the nearby points in the evaluated distribution. When the evaluated distribution contains noise, a greater range of dose values is made available for comparison with a given reference point. Therefore, the likelihood of finding a reference point that satisfies the dose and distance criteria increases. When the roles of the two distributions are reversed, the opposite is true, as extreme dose values from the film distribution are compared to a narrow range of nearby dose values in the smooth Eclipse distribution.

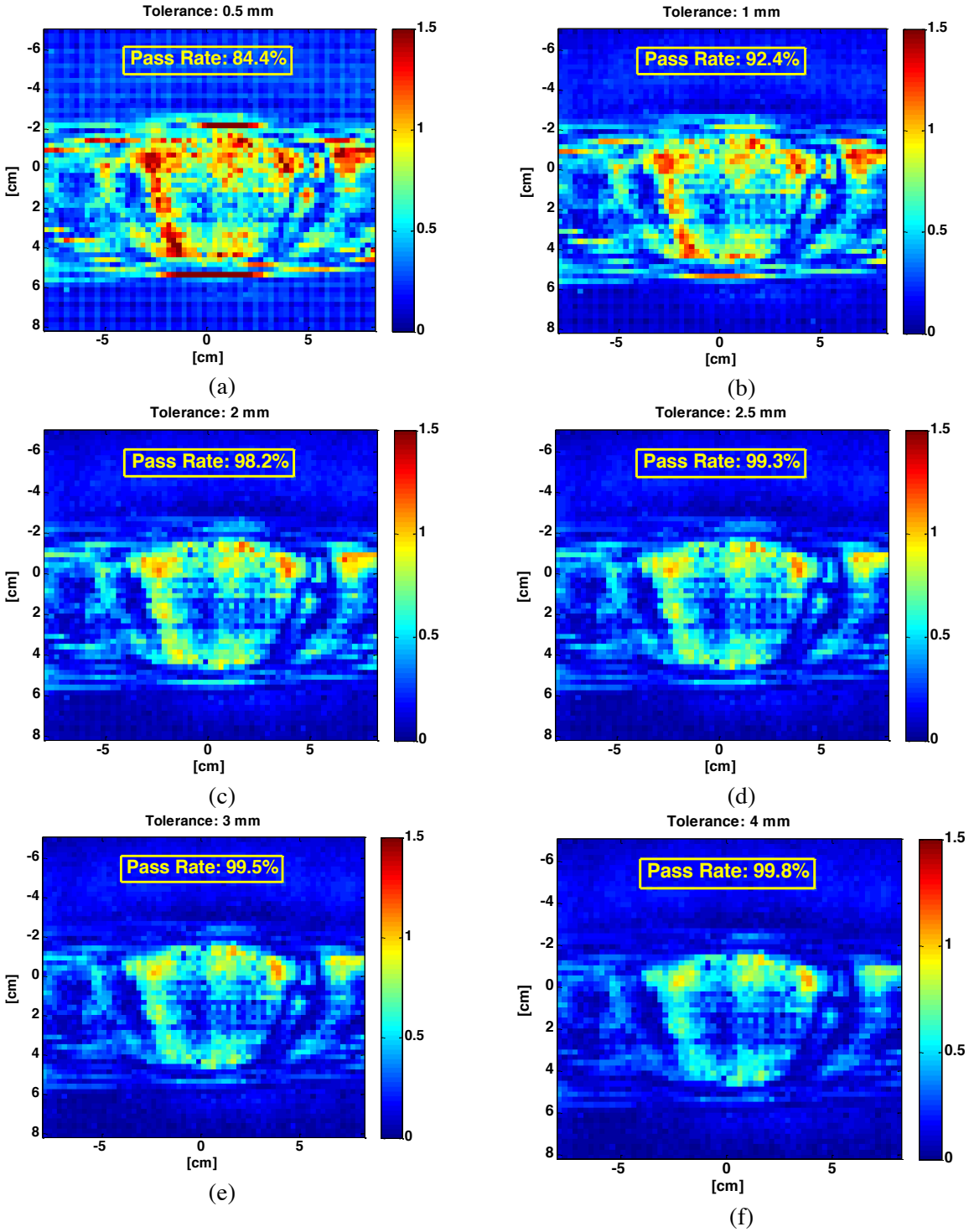

Figure 5: γ maps comparing the 2.5 mm Eclipse data from figure 3 as the reference distribution and the 0.24 mm film data as the evaluated distribution with the γ comparisons performed with different DTA criteria (labelled tolerance in each panel). Panes (a)-(f) display the results of altering the spatial DTA criterion of the γ test between 0.5 mm and 4 mm. Note that the agreement breaks down as DTA becomes smaller than the inherent resolution of the reference dose distribution.

The discrepancy in pass rate and appearance between the two χ distributions (in figure 4) highlights the effect of noise in the reference distribution. The noisy film measurement, when assigned to the role

of reference distribution produces a more forgiving comparison. This is a result of the high gradient regions introduced by noise in the film measurement. The observed increase in gradient increases the acceptance region used when comparing reference and evaluated points, lowering the χ value at that point. The opposite is true when the smooth evaluated distribution serves as the reference distribution.

7. Other considerations and conclusions

Dose comparison is daily task for many medical physicists and γ and χ comparisons have become a common tool for treatment validation. Unfortunately, some of the characteristics of the dose distribution comparison tools are not well understood, and the evaluations can be misused. For example, the authors have seen manuscripts with γ dose distribution comparisons using DTA criteria smaller than the resolution of the data sets, which can lead to inappropriate γ maps (see figure 5). A common practice of reporting the agreement of planned and measured dose delivery (for example, in the acceptance of a particular patient's planned delivery) using only the percentage of points failing the γ evaluation as an assessment. This can reduce a useful evaluation (especially if mapped over structure sets from the planning process) to a single number which may not reflect the clinical relevance of the failures (for example, if the failure is under-dosing in a target area or over-dosing in an organ at risk).

Perhaps this report will encourage readers to take up further study and analysis of dose comparison techniques to more clearly set standards for use. More investigations to establish improved techniques (for example, by probing γ function behaviour with various dose and DTA acceptance criteria, Δd_M and ΔD_M [21]) and procedures for use would benefit the community.

8. Acknowledgements

Funding enabling our investigations of dose evaluation has been provided by the Canadian Institutes of Health Research (CIHR, FRN# 82914) and the Ontario Consortium for Adaptive Interventions in Oncology (OCAIRO) through the Ontario Research Excellence Program. We thank Dr. Daniel Low for providing test distribution approach from his work for us to use in our development, and Mr. Chris Jechel for his work generating the examples illustrating the nature of the dose evaluations.

9. References

[1] Yin F F *et al* 2010 *J. Phys.: Conf. Ser.* **250** 012002
[2] Schreiner L J 2006 *J. Phys.: Conf. Ser.* **56** 1
[3] O'Daniel J *et al* 2010 *J. Phys.: Conf. Ser.* **250** 012050
[4] Mijnheer B *et al* 2010 *J. Phys.: Conf. Ser.* **250** 012020
[5] Schreiner LJ 2009 *J. Phys.: Conf. Ser.* **164** 012001
[6] Schreiner L J *et al* 2011 Proc. Internat. Symp. Standards, Applications and Quality Assurance in Medical Radiation Dosimetry (IDOS) 2, 197-205 (IAEA Vienna, Austria)
[7] Low D 2010 *J. Phys.: Conf. Ser.* **250** 012071
[8] Harms W B *et al* 1998 *Med. Phys.* **25** 1830-6
[9] Low D A *et al* 1998 *Med. Phys.* **25** 656-61
[10] Mah E *et al* 1989 *Phys. Med. Biol.* **34** 1179-94
[11] Van Dyk J *et al* 1993 *Int. J. Radiat. Oncol. Biol. Phys.* **26** 261-73
[12] Depuydt T *et al* 2002 *Radiother. Oncol.* 309-19
[13] Low D A and Dempsey J F 2003 *Med. Phys.* **30** 2455-64
[14] Bakai A *et al* 2003 *Phys. Med. Biol.* **48** 3543-53
[15] Stock M *et al* 2005 *Phys. Med. Biol.* **50** 399-411
[16] Spezi E, and Lewis D G 2006 *Radiother. Oncol.* **79** 224-30
[17] Wendling M 2007 *Med. Phys.* **34** 1647-54
[18] Ju T *et al* 2008 *Med. Phys.* **35** 879-87
[19] Holmes O *et al* 2012 *J. Phys.: Conf. Ser.* (IC3DDose 2012 Proceedings)
[20] Clasie B M *et al* 2012 *Phys. Med. Biol.* **57** 6981-97
[21] Childress N L *et al* 2005 *Med. Phys.* **32** 838-50