Improving Patient-Specific Pre-Treatment Quality Assurance

by

Heming Zhen

A dissertation submitted in partial fulfillment for the degree of Doctor of Philosophy (Medical Physics)

at

UNIVERSITY OF WISCONSIN - MADISON
2012

Date of final oral examination: 05/15/2012

The dissertation is approved by the following members of the Final Oral Committee:
Wolfgang A. Tomé, PH.D., Professor, Human Oncology and Medical Physics
Bhudatt R. Paliwal, PH.D., Professor, Human Oncology and Medical Physics
Bruce R. Thomadsen, PH.D., Professor, Human Oncology and Medical Physics
Michael W. Kissick, PH.D., Assistant Professor, Human Oncology and Medical Physics
Bryan P. Bednarz, PH.D., Assistant Professor, Human Oncology and Medical Physics
Benjamin E. Nelms, PH.D., Clinical Assistant Professor, Human Oncology
“The emphasis should be on why we do a job.”

W. Edwards Deming
Abstract

Rapid technology innovation in radiation therapy has lead to many advanced treatment techniques. A ‘side effect’ of this development is that they have also introduced new sources of error that sometimes make radiation therapy treatment more fragile. As a result, the role of pre-treatment patient-specific QA in ensuring the quality of treatment and safety of patients has become more important than ever.

This dissertation presents the effort to improve the process of pre-treatment patient-specific dose QA. Conventionally, patient-specific QA is performed by comparing the planned and measured phantom dose, and action levels are based on the Gamma passing rate that is generated from this comparison. We started by exploring the question ‘Is Gamma passing rate correlated to clinically relevant patient dose errors’. Through a correlation study based on a virtual QA simulation scheme, we found only weak to moderate correlation between the planar QA passing rate and patient DVH errors. We then went on to perform similar study on 3D Gamma passing rate, both in phantom and in-patient, and determined the conventional QA lacks predictive power to clinically important patient dose error. Many false negative and false positive cases could be created during conventional Gamma-based QA.

In the second part of this dissertation, we aimed to explore new methods and metrics. We first evaluated a commercially available QA system that is capable of predicting the delivered patient dose from conventional QA measurement. The predicted patient dose was shown to be accurate, which opens the possibility of patient dose prediction based QA. We then explored the use of TCP and NTCP models as metrics for this new QA scheme. Through QA simulation under various types of induced errors, we have shown that ?TCP and ?NTCP are potentially good metrics for patient-specific QA. Using these new metrics will allow one to pass false positives and allow one to concentrate on errors that have potentially a large clinical impact. Through this process we have also demonstrated the potential use of these radiobiological models to evaluate the robustness of treatment plans to perturbation introduced by various sources of error.
Acknowledgements

First and foremost I would like to thank my supervisor, Dr. Wolfgang A. Tomé, without whose guidance this project will never be done. Throughout the past four years Dr. Tomé has been the person who with his knowledge, wisdom, and scientific attitude, has not only helped me through each specific research question, but also impacted me greatly as a researcher in general. At several of my critical decision points, Dr. Tomé was the most trustworthy person I turned to. One simply could not wish for a better advisor.

I would also express my sincere appreciation to all my committee members, Dr. Benjamin E. Nelms, Dr. Bhudatt R. Paliwal, Dr. Bruce R. Thomadsen, Dr. Michael W. Kissick, and Dr. Bryan Bednarz, whose advices and suggestions have helped me much improved this work and turned this project to the correct direction. I would especially like to thank Dr. Nelms, who is the major collaborator of this project. His contagious enthusiasm and energy has seriously impacted both the project and myself.

This project requires a lot of clinical knowledge, and I learned them during the joyful time I spent with the nice people in the human oncology clinics. My appreciation goes to Dr. Tolakanahalli, Dr. Tewatia, Dr. Basavatia, Dr. Bayliss, and Dr. Urban for their help in all the clinical works. Among them I would like to specially thank Dr. Tolakanahalli, for she has taught me the most knowledge and logic in the clinical world, and has been a great leader of Team Tomo.

I also want to thank my colleagues and classmates, Dr. Nick Hardcastle, Dr. Edward Bender, Dr. Dongxu Wang, Jialu Yu, Andy Eliss, Suhong Yu, Chi Ma, Yin Zhang, Ke Li, and Poonam Yadav, who have helped me a lot during my PhD study and also made the past four years much more colorful.

Finally, I owe my deepest thanks to my parents, my beautiful fiancé Yunzhi Lin, and my long-life friends Chuan Qin, Hao Tu, Jiwei Liu, Hui Tang, Hao Zhou, Chunyu Zhang, and Xinru Liu. Their unconditional love and support is the greatest gift a person could ever have.
Contents

Abstract ii
Acknowledgements iii
List of Figures vii
List of Tables ix
Abbreviations x

1 Introduction 1
  1.1 Introduction to IMRT . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1
  1.2 Conventional Quality Assurance (QA) for IMRT . . . . . . . . . . . . . . . . . . . . . . 4
  1.3 Scope of this dissertation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

2 Correlation Study of 2D Gamma vs Patient DVH Error 6
  2.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6
  2.2 Materials and Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9
    2.2.1 Experimental design and data acquisition . . . . . . . . . . . . . . . . . . . . . . . . 9
    2.2.2 Correlation of IMRT QA metrics vs clinical DVH metrics . . . . . . . . . . . . . . . . 12
  2.3 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14
    2.3.1 Correlation of IMRT QA metrics vs clinical DVH metrics . . . . . . . . . . . . . . . . . 14
    2.3.2 Investigation of inverse correlation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23
  2.4 Discussion and Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24

3 Correlation Study of 3D Gamma vs Patient DVH Error 30
  3.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 30
  3.2 Materials and Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 33
    3.2.1 Virtual QA using 3D dosimeters with three distinct detector geometries . . . . . . . . . 35
    3.2.2 Virtual QA using 3D patient dose grids . . . . . . . . . . . . . . . . . . . . . . . . . . . 38
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3</td>
<td>Result</td>
<td>39</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Correlation of 3D phantom Gamma passing rates (three geometries) to DVH-based metrics</td>
<td>39</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Correlation of 3D patient Gamma passing rates (whole volume, per ROI) to DVH-based metrics</td>
<td>45</td>
</tr>
<tr>
<td>3.4</td>
<td>Discussion and Conclusion</td>
<td>53</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Correlation of 3D phantom Gamma passing rates (three geometries) to DVH-based metrics</td>
<td>53</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Correlation of 3D patient Gamma passing rates (whole volume, per ROI) to DVH-based metrics</td>
<td>54</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Notes on Local Percent Difference and 2D Composite QA</td>
<td>56</td>
</tr>
<tr>
<td>4</td>
<td>Evaluation of a Patient Dose Prediction Method</td>
<td>59</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>59</td>
</tr>
<tr>
<td>4.2</td>
<td>Materials and Methods</td>
<td>62</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Virtual QA Scheme with 3DVH</td>
<td>62</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Patient dose/DVH prediction with PDP</td>
<td>63</td>
</tr>
<tr>
<td>4.3</td>
<td>Results</td>
<td>66</td>
</tr>
<tr>
<td>4.4</td>
<td>Discussion and Conclusion</td>
<td>72</td>
</tr>
<tr>
<td>5</td>
<td>The Use of Radiobiological Models in Patient-Specific QA</td>
<td>76</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>76</td>
</tr>
<tr>
<td>5.2</td>
<td>Materials and Methods</td>
<td>80</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Error simulation</td>
<td>80</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Virtual QA scheme</td>
<td>82</td>
</tr>
<tr>
<td>5.2.3</td>
<td>Data analysis</td>
<td>82</td>
</tr>
<tr>
<td>5.3</td>
<td>Results</td>
<td>86</td>
</tr>
<tr>
<td>5.3.1</td>
<td>$\Delta$TCP and $\Delta$NTCP for different error types</td>
<td>86</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Correlation between $\Delta$DVH and $\Delta$TCP&amp;$\Delta$NTCP</td>
<td>98</td>
</tr>
<tr>
<td>5.4</td>
<td>Discussion</td>
<td>101</td>
</tr>
<tr>
<td>5.4.1</td>
<td>The use of $\Delta$TCP and $\Delta$NTCP as patient-specific QA metric</td>
<td>101</td>
</tr>
<tr>
<td>5.4.2</td>
<td>$\Delta$TCP and $\Delta$NTCP for different error types</td>
<td>102</td>
</tr>
<tr>
<td>5.4.3</td>
<td>Indication on pre-treatment assessment of plan robustness</td>
<td>104</td>
</tr>
<tr>
<td>5.4.4</td>
<td>Correlation between $\Delta$DVH and $\Delta$TCP&amp;$\Delta$NTCP</td>
<td>105</td>
</tr>
<tr>
<td>6</td>
<td>Conclusion and Future Work</td>
<td>108</td>
</tr>
<tr>
<td>6.1</td>
<td>Is What We’re Doing Adequate?</td>
<td>108</td>
</tr>
<tr>
<td>6.2</td>
<td>Is It Possible to Change?</td>
<td>109</td>
</tr>
<tr>
<td>6.3</td>
<td>What’s Next?</td>
<td>110</td>
</tr>
</tbody>
</table>
## List of Figures

2.1 2D Gamma correlation study scheme ........................................ 10
2.2 Sample error induced DVH .................................................. 15
2.3 Frequency distribution of simulated DVH error ............................ 16
2.4 CTV D95 error vs 2D Gamma ................................................. 17
2.5 Cord D1cc error vs 2D Gamma ............................................... 17
2.6 Cord max dose error vs 2D Gamma ......................................... 18
2.7 Contralateral parotid mean dose error vs 2D Gamma .................... 18
2.8 Ipsilateral parotid mean dose error vs 2D Gamma ....................... 19
2.9 Larynx mean dose error vs 2D Gamma .................................... 19
2.10 CTV D95 error vs 2D Gamma (by error type) ............................. 25
2.11 Illustration of false negatives and false positives ....................... 27

3.1 3D Gamma correlation study scheme ....................................... 34
3.2 Illustration of three different 3D phantom geometries .................. 37
3.3 CTV D95 error vs 3D phantom Gamma .................................... 40
3.4 Spinal cord D1cc error vs 3D phantom Gamma ........................... 40
3.5 Spinal cord max dose error vs 3D phantom Gamma ..................... 41
3.6 Contralateral parotid mean dose error vs 3D phantom Gamma ........ 41
3.7 Ipsilateral parotid mean dose error vs 3D phantom Gamma .......... 42
3.8 Larynx mean dose error vs 3D phantom Gamma ........................ 42
3.9 CTV D95 error vs 3D patient Gamma ..................................... 46
3.10 Spinal cord D1cc error vs 3D patient Gamma ............................ 47
3.11 Spinal cord max dose error vs 3D patient Gamma ..................... 48
3.12 Contralateral parotid mean dose error vs 3D patient Gamma ........ 49
3.13 Ipsilateral parotid mean dose error vs 3D patient Gamma .......... 50
3.14 Larynx mean dose error vs 3D patient Gamma ........................ 51
3.15 Example of local percent difference and planar composite QA ....... 58

4.1 Actual vs PDP-predicted patient DVH error for CTV D95 .............. 67
4.2 Actual vs PDP-predicted patient DVH error for spinal cord D1cc ... 67
4.3 Actual vs PDP-predicted patient DVH error for spinal cord Dmax .... 68
4.4 Actual vs PDP-predicted patient DVH error for contralateral parotid mean dose ......................................................... 68
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.5</td>
<td>Actual vs PDP-predicted patient DVH error for ipsilateral parotid mean dose</td>
</tr>
<tr>
<td>4.6</td>
<td>Actual vs PDP-predicted patient DVH error for larynx mean dose</td>
</tr>
<tr>
<td>4.7</td>
<td>PDP prediction accuracy boxplot</td>
</tr>
<tr>
<td>4.8</td>
<td>Histogram of dose error before and after PDP correction</td>
</tr>
<tr>
<td>5.1</td>
<td>Sigmoid shaped TCP and NTCP curve</td>
</tr>
<tr>
<td>5.2</td>
<td>Matlab module for TCP/NTCP calculation</td>
</tr>
<tr>
<td>5.3</td>
<td>H&amp;N CTV DVH curves</td>
</tr>
<tr>
<td>5.4</td>
<td>H&amp;N GTV DVH curves</td>
</tr>
<tr>
<td>5.5</td>
<td>Contralateral parotid DVH curves</td>
</tr>
<tr>
<td>5.6</td>
<td>Ipsilateral parotid DVH curves</td>
</tr>
<tr>
<td>5.7</td>
<td>Spinal cord DVH curves</td>
</tr>
<tr>
<td>5.8</td>
<td>Larynx DVH curves</td>
</tr>
<tr>
<td>5.9</td>
<td>Prostate CTV DVH curves</td>
</tr>
<tr>
<td>5.10</td>
<td>Rectal wall DVH curves</td>
</tr>
<tr>
<td>5.11</td>
<td>Range of TCP for H&amp;N CTV</td>
</tr>
<tr>
<td>5.12</td>
<td>Range of TCP for H&amp;N GTV</td>
</tr>
<tr>
<td>5.13</td>
<td>Range of NTCP for contralateral parotid</td>
</tr>
<tr>
<td>5.14</td>
<td>Range of NTCP for ipsilateral parotid</td>
</tr>
<tr>
<td>5.15</td>
<td>Range of NTCP for spinal cord</td>
</tr>
<tr>
<td>5.16</td>
<td>Range of NTCP for larynx</td>
</tr>
<tr>
<td>5.17</td>
<td>Range of TCP for prostate CTV</td>
</tr>
<tr>
<td>5.18</td>
<td>Range of NTCP for rectal wall</td>
</tr>
<tr>
<td>5.19</td>
<td>TCP change vs CTV D95 change for prostate cases</td>
</tr>
<tr>
<td>5.20</td>
<td>NTCP change vs change in ipsilateral parotid mean dose</td>
</tr>
<tr>
<td>5.21</td>
<td>NTCP change vs change in larynx mean dose</td>
</tr>
<tr>
<td>5.22</td>
<td>Different regions on TCP curve</td>
</tr>
<tr>
<td>5.23</td>
<td>Different regions on NTCP curve</td>
</tr>
</tbody>
</table>
List of Tables

1.1 IMRT error source .................................................. 3

2.1 Correlation of DVH error to 2D Gamma .......................... 21
2.2 DVH error with high Gamma passing rate ....................... 22
2.3 Correlation of DVH error to 2D Gamma (by error types) ....... 24

3.1 Correlation of patient DVH error to 3D phantom Gamma ...... 44
3.2 Correlation of patient DVH error to 3D patient Gamma ........ 52

4.1 Correlation of predicted vs actual patient DVH error ........... 70

5.1 Change in H&N TCP and NTCP by error type ................... 96
5.2 Change in prostate TCP and NTCP by error type ............... 97
5.3 Correlation between change in TCP/NTCP and DVH points .... 99
Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMRT</td>
<td>Intensity Modulated Radiation Therapy</td>
</tr>
<tr>
<td>QA</td>
<td>Quality Assurance</td>
</tr>
<tr>
<td>QM</td>
<td>Quality Management</td>
</tr>
<tr>
<td>TPS</td>
<td>Treatment Planning System</td>
</tr>
<tr>
<td>MLC</td>
<td>Multi Leaf Collimator</td>
</tr>
<tr>
<td>DVH</td>
<td>Dose Volume Histogram</td>
</tr>
<tr>
<td>MU</td>
<td>Monitor Unit</td>
</tr>
<tr>
<td>DTA</td>
<td>Distance To Agreement</td>
</tr>
<tr>
<td>OAR</td>
<td>Organ At Risk</td>
</tr>
<tr>
<td>ROI</td>
<td>Region Of Interest</td>
</tr>
<tr>
<td>PTV</td>
<td>Planning Target Volume</td>
</tr>
<tr>
<td>CTV</td>
<td>Clinical Target Volume</td>
</tr>
<tr>
<td>GTV</td>
<td>Gross Tumor Volume</td>
</tr>
<tr>
<td>H&amp;N</td>
<td>Head and Neck</td>
</tr>
<tr>
<td>T&amp;G</td>
<td>Tongue and Groove</td>
</tr>
<tr>
<td>TCP</td>
<td>Tumor Control Probability</td>
</tr>
<tr>
<td>gEUD</td>
<td>generalized Equivalent Uniform Dose</td>
</tr>
<tr>
<td>NTCP</td>
<td>Normal Tissue Complication Probability</td>
</tr>
<tr>
<td>UTCP</td>
<td>Uncomplicated Tumor Control Probability</td>
</tr>
</tbody>
</table>

Dedicated to my farther, Hongying Zhen, and my mother, Keping Zhang, for all their love that makes me become who I am.

谨以此文献给我的父亲甄鸿英和母亲张克平,他们的爱使我成为了今天的我。
Chapter 1

Introduction

1.1 Introduction to IMRT

The goal of radiation therapy is to deliver high radiation dose to the tumor while sparing normal tissue to avoid possible side effect. A significant leap during the past two decades in pursuing this goal is the innovation of Intensity Modulated Radiation Therapy (IMRT). Instead of using blocks and wedges to shape the beam, multi-leaf collimator (MLC) was implemented to modify both the shape of the beams and the intensity distribution within each beam. These tungsten leaves can move smoothly in and out of the treatment field within a very short duration of time. The motion of the MLC is driven by a motor and controlled by computers, which has made the IMRT process less labor intensive and less time consuming.
The appearance of IMRT has also changed the way treatment planning is carried out. Before IMRT, radiation treatment employed a ‘forward planning’ process, during which physicist/oncologist manually set the beams, created modifiers, and then calculated the dose distribution. IMRT utilizes a deconvolution technique called ‘inverse planning’ that automates this process with an optimization computer program. In IMRT treatment planning, the intensity matrix of each beam instead of the machine parameters are often optimized using a cost function containing physical and/or biological dose constraints. For faster computation in this iterative process, assumptions of homogeneous material and no scatter are often adopted, as well as the technique of random sampling dose volume histogram (DVH). Once an optimal intensity matrix is found, a secondary computation is performed to determine the sequence of field shape and associated monitor units (MUs) to navigate the MLC motion. Final dose distribution is calculated using a more accurate superposition/convolution algorithm[1] which considers scatter and inhomogeneity.

With the development of the MLC and inverse planning technique, IMRT has the ability to create a much more complex dose distribution than its predecessor – three dimensional conformal therapy (3D-CRT). Dose distributions can be accurately planned to have steep gradient and concave shapes that wraps around organs at risk (OAR), whose dose limits are significantly lower than the tumor dose. The implementation of IMRT in clinic has been growing rapidly since it was invented. In fact, it was reported in the year 2008 that 30%-60% cancer patients are treated by IMRT[2].

It is important to be aware that the whole process of IMRT is so complex that various
types of error could occur. Also, with the highly conformal dose distribution created by IMRT, errors and uncertainties in the treatment process may have bigger effects than before. Errors in target/OAR contouring, patient setup and MLC motion could easily cause cold spots in the target and/or overdose in the critical structures. Many studies have been conducted in the effort of addressing, quantifying and reducing these uncertainties and errors in IMRT[3–7]. A brief summary of possible error sources in IMRT is shown in Table 1.1. Several specific types of errors among them will be studied throughout this thesis work. For the safety of patients, extreme caution is required in dealing with these errors in clinical practice. A recent paper has reported several clinical incidents happened in the US and Europe in which patients were overdosed due to errors such as incorrect machine calibration and MLC file corruption during plan transfer[7].

<table>
<thead>
<tr>
<th>Category</th>
<th>Possible error/uncertainty source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commissioning phase</td>
<td>Inaccurate beam model (penumbra, MLC transmission, leaf offset, tongue-and-groove effect, etc.)</td>
</tr>
<tr>
<td>Planning Phase</td>
<td>Planning image quality (distortion, artifact), IVDT, target and organ contouring uncertainty, dose calculation algorithm limitation</td>
</tr>
<tr>
<td>Transferring phase</td>
<td>Error Image, plan, structure and MLC sequence information transfer</td>
</tr>
<tr>
<td>Delivery phase</td>
<td>Patient setup error, inter &amp; intra fraction organ motion, machine output drifting</td>
</tr>
<tr>
<td>Human error</td>
<td>Unexpected human error can happen in any phase</td>
</tr>
</tbody>
</table>
1.2 Conventional Quality Assurance (QA) for IMRT

The complexity of IMRT (as well as other treatment techniques) necessitates dedicated QA programs to ensure the quality of treatment and safety of patients. In general, clinics employ two types of QA programs: a machine QA program and a patient-specific QA program. Machine QA programs check and demonstrate that the machine characteristics do not deviate significantly from their baseline values acquired at the time of acceptance and commissioning.[8] Depending on the reliability and importance of each specific machine functionality (such as output, MLC position, couch, gantry and jaw motion and beam quality), these QA tests are performed repeatedly at different frequencies (daily, weekly, monthly and annually). Several task group (TG) reports from the American Association of Physicists in Medicine (AAPM) have been generated to provide guidelines for such machine QA programs.[9–11]. It is worth noting that all these tests are machine-specific rather than patient-related. The principle in setting tolerance of each test is that the combined uncertainty of all components should not cause more than ±5% error for a single dose point.[11]

In addition to the machine QA program, a patient-specific QA program is also in place to ensure the quality of each individual patient treatment (especially for advanced treatment techniques such as IMRT, SBRT and arc therapy). The main purpose of patient-specific dose QA should not be to assure each specific functionality of the machine, but to assure that the clinical impact of the treatment on the patient, due to the overall performance of the machine and all human factors, does not deviate significantly from what
is planned. Conventionally, physicists perform patient-specific IMRT QA by applying the patient dose to a 2D/3D phantom and comparing the planned and measured phantom dose[12, 13]. Clinical decisions are made based on result of such comparison, which is conventionally quantified using the Gamma metric.[14] This conventional method and metric will be further described in Chapter 2.

1.3 Scope of this dissertation

This dissertation is focused on the method and metric for patient-specific dose QA. In Chapter 2 we will investigate, through a QA simulation, the correlation between clinically important patient DVH errors and per-beam Gamma passing rate in planar IMRT QA – the most common IMRT QA scheme – in order to determine whether 2D Gamma passing rates predict patient DVH errors. We will generalize this study to 3D detector designs in Chapter 3, and furthermore explore Gamma passing rates in the patient volume. This will allow us to determine the predictive power of 3D Gamma, and discuss whether the current QA method and metric are intrinsically sufficient for the purpose of patient-specific dose QA. In Chapter 4, we will turn to a new patient-specific QA method that is based on patient dose prediction, and use our simulated QA data to evaluate a commercial software that utilizes such QA approach. In Chapter 5, we will explore the possibility of new metrics based on the QA method studied in Chapter 4. Specifically, we will investigate the use of radiobiological models in patient-specific dose QA.
Chapter 2

Correlation Study of 2D Gamma vs Patient DVH Error

2.1 Introduction

In modern radiation therapy, each patient treatment plan is customized and unique. In the case of IMRT, each treatment field can be highly complex and justifies patient-specific QA to verify (1) the treatment planning system’s ability to calculate the dose accurately and (2) the delivery system’s ability to deliver the dose accurately. A description of the conventional IMRT patient specific QA can be found in the 2003 ‘Guidance Document’ of IMRT[13], “Verification measurements are commonly made of a ‘phantom plan’ or ‘hybrid plan.’” This technique consists of applying the MLC segments, leaf trajectories and MU for each field, derived from the final patient calculation, to a CT study
of a standard phantom and then recalculating the final deliverable dose distribution in
the phantom.” The phantom is then irradiated according to this plan and the doses
measured using ion chambers, film, or other detectors. The results of the measurement
are then compared to the predicted dose to the phantom. The logic behind this method-
ology is that if the phantom measurement agrees with the planned on a certain level, it
can be safely assumed that the plan is deliverable and patient dose will agree with the
plan on a certain level.

For QA measurement, an ion chamber is generally used for absolute dose verification
of a certain dose point, and a film can be used to acquire a high resolution relative
dose distribution. Alternatively, a 2-dimensional detector array, such as an ion chamber
array[15–19] or a diode array[20, 21] can be used for absolute dose distribution verifi-
cation. The application of these array detectors has two immediate benefits: they are
much less time-consuming and less labor-intensive because their instant readouts allow
QA analysis to be done right after the delivery, while for film-based QA it is often neces-
sary to wait several hours before the film is stable to be processed. Also, no pre-analysis
processing (such as film developing) is needed, which reduces the sources of error. How-
ever, some other issues, including non-uniform angular dependence[22] and insufficient
resolution (due to the sparse density) arise that need further consideration.

Gamma analysis[14] is the most commonly used method for comparing the planned and
measured dose distribution. For each pixel in the dose distribution, a gamma index is
calculated based on the normalized percent dose difference and distance to agreement
(DTA). A Gamma index that is smaller than or equal to unity will usually be considered
a ‘pass’ for the specific pixel. Calculating the Gamma indices for all pixels gives the Gamma passing rate — the percentage of pixels that have Gamma indices smaller than or equal to unity. Gamma (or DTA) passing rate is the most commonly used metric based on which QA decisions are made. In the recent report of the AAPM Task Group 119 (TG119)[12] on IMRT commissioning, the ‘3%/3mm’ (normalization criteria for percent dose difference and DTA) Gamma passing rate is chosen as the IMRT QA metric. In fact, a recent report[23] shows that ‘3%/3mm’ Gamma passing rate is the most commonly used QA metric by clinicians.

There has not been a well-defined standard on when (at what passing rate) to approve a QA. Several studies have been performed aiming to suggest acceptance/action levels for planar IMRT QA[12, 24–28], and most of these studies base the suggested action levels on retrospective statistical analysis of the performance that have been achieved over many plans and IMRT beams. It has been suggested that meeting such action levels should be a requirement in order to take part in clinical trials[24]. However, what is of ultimate importance in patient-specific QA is the impact on the treatment outcome, often estimated based on the patient DVH values. In fact, no substantial study has been done to determine whether Gamma passing rate is an appropriate metric for patient-specific dose QA. Before setting any action level based on Gamma passing rate, it needs to be made clear whether achieving a high (low) Gamma passing rate indicates small (big) patient DVH error. In other words, it has yet to be proven or disproved whether there is a strong correlation between Gamma passing rate and clinically important patient DVH errors.
The purpose of this chapter is to explore the statistical correlation of conventional IMRT QA performance metrics to per-patient/plan clinically relevant dose difference metrics and, in the process, determine if today’s standards and published action levels (which have been based on the statistics of what is commonly achieved) are justified.

2.2 Materials and Methods

2.2.1 Experimental design and data acquisition

Twenty-four clinically approved and treated head and neck IMRT treatment plans were chosen from our database and fully anonymized for the purpose of this study. All plans were generated using the Pinnacle (Philips Radiation Oncology Systems, Fitchburg, WI) TPS (Treatment Planning System) using 6 MV x-ray beams from Varian (Palo Alto, CA) linear accelerators with 120-leaf MLC. Head and neck IMRT plans are highly complex, with multiple target volumes as well as multiple organs at risk distributed throughout the treatment volume. Hence, these plans were desirable for QA sensitivity and specificity analysis.

A schematic of our simulation methodology is summarized in Fig 2.1. It is important to note that inducing beam model errors serves the purpose of creating dose differences that can be simulated and quantified in both the planar IMRT QA schema as well as in the patient model. The purpose was not to study beam model errors per se, but rather to create a system where IMRT QA metrics and patient anatomy-based dose differences could be quantified and used to study statistical correlation. In all cases, the error-free
beam model was used as the “simulated measurement” for the IMRT QA planar dose and patient dose. The degree of the errors induced for this study were selected to result in realistic IMRT QA performance metrics, i.e., passing rates commonly accepted in clinical practice without further investigation.

In order to simulate errors with impact on dose gradients and dose levels, we generated four experimental beam models. Two beam models were modified to calculate a shallower penumbra than the error-free model, while the other two beam models were modified by (1) halving and (2) doubling the MLC transmission of the error-free model. For the first two beam models, hereafter called the shallow penumbra beam model
(SPBM) and the very shallow penumbra beam model (VSPBM), we modified the error-free penumbra (80%–20%) of 4.5 \(d_{\text{max}}\) and 5.9 mm (depth 10 cm) for a \(10 \times 10 \text{ cm}^2\) open field in a water phantom, calculated on a \(1 \text{ mm} \times 1 \text{ mm} \times 1 \text{ mm}\) dose grid. The modified SPBM penumbra (80%–20%) was 7.2 \(d_{\text{max}}\) and 9.2 mm (depth 10 cm). The modified VSPBM penumbra (80%–20%) was 8.6 \(d_{\text{max}}\) and 11.0 mm (depth 10 cm). The third experimental model was the high transmission beam model, for which the error-free MLC transmission (1.94%) was doubled (3.88%). The final experimental model was the low transmission beam model, for which the MLC transmission was halved (0.97%).

For each of the 24 head and neck patients, four new IMRT plans were generated using each of the modified beam models that have been described above and each employing the same dose objectives and number of iterations. All 3D patient plans were calculated on a \(4 \text{ mm} \times 4 \text{ mm} \times 4 \text{ mm}\) dose grid. QA dose planes were calculated with the following parameters: \(1 \text{ mm} \times 1 \text{ mm}\) resolution, normal to the beam axis at source-to-plane distance 100 cm, depth 10 cm, in a flat homogeneous phantom, and using the patient plan beam MU. (Note: Using this method simulates a full density and high resolution IMRT QA plane, i.e., equivalent to an ideal film and not the sparse density of commercial arrays made of diodes or ion chambers.) Then, the same 96 plan doses were recalculated using the error-free beam model but with all other parameters held constant (i.e., all beam parameters, IMRT segment shapes and weights, and the monitor units for each segment/beam were set equal to those arrived at in the original optimization with the modified beam models). Thus, the only sources of variation between the 96 pairs of plans were the beam model modifications. The following data were exported from the TPS
per-patient plan: (1) DICOM RT plan, (2) DICOM RT structure set, (3) 3D patient dose volume as DICOM RT dose, and (4) 2D dose planes as ASCII text files per beam.

In this study, the error-free beam model was used to produce virtual measurements on the virtual linear accelerator. This allows for a controlled study since it eliminates output variations present in a real medical linear accelerator and allows one to compare planar IMRT QA films of optimal data density, independent of any density or resolution limitations of commercial arrays.

### 2.2.2 Correlation of IMRT QA metrics vs clinical DVH metrics

The “simulated-measured” and planned QA planar dose planes were analyzed using MAPCHECK software (Sun Nuclear Corporation) employing the Gamma passing rate metric, which is a common metric employed in conventional IMRT QA. QA scores (percentage of dose points with a gamma value less than 1) were generated for each pair of planes using the following Gamma criteria: 1%/1 mm, 2%/2 mm, and 3%/3 mm, where the percent is the per-voxel dose difference given as a percent of global normalization dose and the distance is the distance-to-agreement criterion. Dose values below 10% of the per-beam normalization (max) dose were ignored. 3DVH software (Sun Nuclear Corporation), an IMRT QA software module capable of quantifying 3D dose comparisons, was used to generate the following anatomy dose metrics for select volumes: Spinal cord max dose, spinal cord dose to 1 cc (D1cc), contralateral parotid mean dose, ipsilateral parotid mean dose, larynx mean dose, and CTV dose to 95% volume (D95). These anatomy dose metrics were generated for both the planned and the simulated-measured
patient dose. The resulting absolute values of the errors of the clinical dose metrics were plotted vs IMRT QA performance metrics (Gamma passing rates as %). The term dose metric “error” is used throughout this report to quantify the difference between the actual dose (generated using the error-free system) and the expected/planned dose (generated using the error-induced system) relative to the expected/planned dose. The dose errors are thus calculated according to the following equations:

\[
\text{Dose Error} \ (\%) = \left( \frac{\text{Actual Dose Value (Gy)} - \text{Planned Dose Value (Gy)}}{\text{Planned Dose Value (Gy)}} \right) \times 100\%,
\]

\[
\text{Absolute Dose Error} \ (\%) = |\text{Dose Error} \ (\%)| \tag{2.1}
\]

To assess correlation, simple linear regression lines and their corresponding Pearson product moment correlation values, hereafter simply denoted as Pearson’s r-values, were generated. In order to quantify the incidence of false negatives, the ranges of the observed clinical dose metric errors along with the average absolute errors were generated for populations of IMRT QA performance metrics with 95+% conventional QA passing rates using 3%/3 mm, 2%/2 mm, and 1%/1 mm Gamma.
2.3 Results

2.3.1 Correlation of IMRT QA metrics vs clinical DVH metrics

Fig 2.2 shows, for the sake of example, the DVH variation between the planned (error-induced beam model) and ‘virtually measured’ (error-free beam model) patient dose in one of the 24 patients. It can be seen that the MLC transmission errors result in a systematic shift of the DVH curves, while the penumbra errors perturb the DVH in a non-uniform way. To demonstrate the impact of different errors on specific DVH endpoints, Fig 2.3 plots the distribution of patient dose errors due to different induced error types for two of the critical anatomical dose metrics (CTV D95 and contralateral parotid mean dose, chosen because these exhibited a small range and a large range of errors, respectively).

Figs. 2.4-2.9 show the magnitude of the anatomy-based dose difference metrics (ordinate) vs conventional IMRT QA passing rates (3%/3 mm, 2%/2 mm, and 1%/1 mm Gamma, global percent normalization to field max, 10% lower threshold cutoff). For each point, the average IMRT QA passing rate (of all fields in each plan) was used as the abscissa value. The r-values are shown in Table 2.1 along with the respective p-values. Table 2.2 gives the ranges and the sample standard deviations of the clinical dose errors for plans exceeding 95% passing rates for two sets of Gamma analyses (3%/3 mm and 2%/2 mm) and exceeding 90% passing rates for Gamma analysis at 1%/1 mm.
Figure 2.2: Sample DVH differences between the induced-error beam models (dashed lines) and the virtual measurement beam models (solid lines). These are the results for patient plan no. 22 (of 24).
Figure 2.3: Distribution of errors in two critical anatomy dose metrics for three types of errors induced. (a) Errors in CTV D95 (low range of errors, overall) and (b) errors in contralateral parotid mean dose (higher range of errors).
Figure 2.4: Magnitude of errors in CTV D95 dose vs the conventional IMRT QA performance metric of passing rate (%) averaged over all beams per plan, shown for three different sets of Gamma parameters.

Figure 2.5: Magnitude of errors in spinal cord D1cc dose vs the conventional IMRT QA performance metric of passing rate (%) averaged over all beams per plan, shown for three different sets of Gamma parameters.
Figure 2.6: Magnitude of errors in spinal cord max dose vs the conventional IMRT QA performance metric of passing rate (%) averaged over all beams per plan, shown for three different sets of Gamma parameters.

Figure 2.7: Magnitude of errors in contralateral parotid mean dose vs the conventional IMRT QA performance metric of passing rate (%) averaged over all beams per plan, shown for three different sets of Gamma parameters.
Figure 2.8: Magnitude of errors in ipsilateral parotid mean dose vs the conventional IMRT QA performance metric of passing rate (%) averaged over all beams per plan, shown for three different sets of Gamma parameters.

Figure 2.9: Magnitude of errors in larynx mean dose vs the conventional IMRT QA performance metric of passing rate (%) averaged over all beams per plan, shown for three different sets of Gamma parameters.
These data clearly show that there are only weak to moderate correlations between conventional IMRT QA performance metrics and anatomy-based dose difference metrics, as evidenced by their corresponding Pearson’s r-values (cf. Table 2.1). Moreover, all moderate correlations ($0.3 < |r| < 0.7$) with statistically significant p-values ($p < 0.05$) have positive Pearson’s r-values, indicating that the larger clinical errors happen for higher IMRT QA Gamma passing rates. This suggests a large incidence of false negatives. As evidenced in Figs. 2.4-2.9 and Table 2.2, some of the largest errors to critical structures occurred when IMRT QA results were very high (95% or above).
Table 2.1: Pearson correlation values (r) and two-tailed p-values correlating the magnitude of anatomy dose errors to three IMRT QA Gamma passing rate performance metrics. (Note: The statistically significant correlations have positive r-value (positive slope) indicating that the highest critical dose errors happen at the higher Gamma passing rates.)

<table>
<thead>
<tr>
<th>IMRT QA criteria*</th>
<th>Spinal cord D1cc error r</th>
<th>p-value</th>
<th>Contralateral parotid mean dose error r</th>
<th>p-value</th>
<th>Ipsilaterial parotid mean dose error r</th>
<th>p-value</th>
<th>Larynx mean dose error r</th>
<th>p-value</th>
<th>CTV D95 error r</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>3%/3 mm</td>
<td>-0.183</td>
<td>0.07</td>
<td>0.328</td>
<td>&lt;0.01</td>
<td>0.523</td>
<td>&lt;0.01</td>
<td>-0.167</td>
<td>0.20</td>
<td>0.604</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>2%/2 mm</td>
<td>-0.141</td>
<td>0.17</td>
<td>0.118</td>
<td>0.25</td>
<td>0.588</td>
<td>&lt;0.01</td>
<td>-0.134</td>
<td>0.31</td>
<td>0.653</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>1%/1 mm</td>
<td>-0.130</td>
<td>0.21</td>
<td>0.10</td>
<td>0.33</td>
<td>0.551</td>
<td>&lt;0.01</td>
<td>-0.295</td>
<td>0.02</td>
<td>0.619</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>

*Analysis criteria method: Global % difference (normalized to max dose), 10% lower threshold, and γ index ≤1 as the passing criterion.
Table 2.2: Range of errors (%) and mean absolute errors (%) for clinically relevant metrics in the case of all plans (N) meeting a specified threshold Gamma passing rate for three sets of Gamma parameters

<table>
<thead>
<tr>
<th>Anatomy dose metric</th>
<th>Range of % Errors</th>
<th>3%/3 mm (N=83)</th>
<th>2%/2 mm (N=51)</th>
<th>1%/1 mm (N=12)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spinal cord</td>
<td>3%/3 mm (N=83)</td>
<td>-11.1, 15.7</td>
<td>-11.1, 15.7</td>
<td>-2.7, 3.3</td>
</tr>
<tr>
<td>D1cc</td>
<td>Mean absolute error (%)</td>
<td>3.2</td>
<td>3.4</td>
<td>2.3</td>
</tr>
<tr>
<td>Contralateral</td>
<td>Range of % Errors</td>
<td>-10.9, 12.0</td>
<td>-10.9, 12.0</td>
<td>-5.1, 5.7</td>
</tr>
<tr>
<td>Parotid mean</td>
<td>Mean absolute error (%)</td>
<td>4.5</td>
<td>5.5</td>
<td>4.0</td>
</tr>
<tr>
<td>Ipsilaterall</td>
<td>Range of % Errors</td>
<td>-3.7, 4.1</td>
<td>-3.7, 4.1</td>
<td>-1.4, 1.7</td>
</tr>
<tr>
<td>Parotid mean</td>
<td>Mean absolute error (%)</td>
<td>1.50</td>
<td>2.1</td>
<td>1.5</td>
</tr>
<tr>
<td>Larynx mean</td>
<td>Range of % Errors</td>
<td>-15.9, 9.2</td>
<td>-7.6, 9.2</td>
<td>-3.2, 3.7</td>
</tr>
<tr>
<td>CTV D95</td>
<td>Mean absolute error (%)</td>
<td>5.7</td>
<td>5.3</td>
<td>2.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Anatomy dose metric</th>
<th>Range of % Errors</th>
<th>3%/3 mm (N=83)</th>
<th>2%/2 mm (N=51)</th>
<th>1%/1 mm (N=12)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean absolute error (%)</td>
<td>-3.7, 2.6</td>
<td>-2.2, 2.6</td>
<td>-1.6, 1.6</td>
</tr>
<tr>
<td></td>
<td>Mean absolute error (%)</td>
<td>1.3</td>
<td>1.7</td>
<td>1.3</td>
</tr>
</tbody>
</table>

Error ranges and the mean absolute errors are given as percent errors (%) using the error-free plans as the baseline.

Analysis criteria method: Global % difference (normalized to max dose), 10% lower threshold, and $\gamma$ index $\leq 1$ as the passing criterion.

The average of error magnitudes, i.e., absolute values of errors (%).

---

a Error ranges and the mean absolute errors are given as percent errors (%) using the error-free plans as the baseline.

b Analysis criteria method: Global % difference (normalized to max dose), 10% lower threshold, and $\gamma$ index $\leq 1$ as the passing criterion.

c The average of error magnitudes, i.e., absolute values of errors (%).
It can be concluded from the result that conventional IMRT QA metrics do not necessarily predict the likelihood of clinically relevant dose difference metrics. Significant errors that could lead to an induction of unwanted normal tissue toxicities (if the dose is pushed above currently acceptable dose tolerance values) or to a decrease in expected local tumor control (+14.8% max cord dose error, +12.0% mean parotid dose error, +9.2% mean larynx dose error, and -3.7% CTV D95 error) happen even at the highest levels of conventional IMRT QA passing rate.

2.3.2 Investigation of inverse correlation

As seen from Table 2.1, both CTV D95 error and ipsilateral parotid mean dose error demonstrate a moderate positive correlation to Gamma passing rate with statistical significance. Such result is counter-intuitive at the first glance since it indicates higher passing rates correlating to bigger DVH errors. In order to seek an explanation of this result, the correlation analysis for CTV D95 and Ipsilateral parotid mean dose was re-conducted with data been separated according to the type of induced errors. Specifically, the data was categorized into two subgroups: MLC transmission error group and MLC penumbra error group, then the correlation coefficients of DVH error to Gamma passing rates, as well as their corresponding p-values were generated for each group of data respectively, as shown in Table 2.3.

These results show that when separated by the types of induced errors, no moderate positive correlation was observed. In fact, contrary to the result from Table 2.1, moderate negative correlation was found for CTV D95 vs Gamma passing rate with statistical
significance. In most other cases (with the exception of Ipsilateral parotid mean dose to 1%/1 mm Gamma passing rate), correlation was found to be weak and has no statistical significance. Figure 2.10 demonstrates, for the sake of example, the scatter plot of CTV D95 error vs Gamma passing rate at 1%/1 mm. It is clearly seen that the subgroup with MLC transmission error demonstrates a negative correlation, while the subgroup of MLC penumbra error demonstrates a very weak positive (not significant) correlation, however when all data of different induced-errors were combined together, inter-group correlation dominates intra-group correlation and the undesired moderate positive correlation appears. It can be concluded that the moderate positive correlation is an artifact due to different characteristics of the induced errors. However, no strong desired correlation was found even within each group, as shown in Table 2.3.

<table>
<thead>
<tr>
<th>Induced Error Type</th>
<th>3%/3 mm r</th>
<th>3%/3 mm p-value</th>
<th>2%/2mm r</th>
<th>2%/2mm p-value</th>
<th>1%/1mm r</th>
<th>1%/1mm p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTV D95 MLC Transmission</td>
<td>-0.225</td>
<td>0.14</td>
<td>-0.476</td>
<td>&lt;0.01</td>
<td>-0.585</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>CTV D95 MLC Penumbra</td>
<td>0.058</td>
<td>0.71</td>
<td>0.08</td>
<td>0.61</td>
<td>0.051</td>
<td>0.74</td>
</tr>
<tr>
<td>Ipsilateral parotid mean dose error MLC Transmission</td>
<td>-0.285</td>
<td>0.09</td>
<td>-0.243</td>
<td>0.15</td>
<td>-0.271</td>
<td>0.11</td>
</tr>
<tr>
<td>Ipsilateral parotid mean dose error MLC Penumbra</td>
<td>-0.104</td>
<td>0.55</td>
<td>-0.23</td>
<td>0.17</td>
<td>-0.51</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>

### 2.4 Discussion and Conclusion

These results show there are only weak to moderate correlations between conventional IMRT QA performance metrics and clinically relevant dose difference metrics, with the
<table>
<thead>
<tr>
<th>Error (% in CTV D95 vs Conventional QA metric) (separated by induced error types)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absolute Error (%) in CTV D95 vs Conventional QA metric</td>
</tr>
<tr>
<td>Gamma Passing Rates (%)</td>
</tr>
<tr>
<td>0.5</td>
</tr>
<tr>
<td>MLC Transmission Error</td>
</tr>
</tbody>
</table>

**Figure 2.10:** Magnitude of errors in CTV D95 vs the conventional IMRT QA performance metric of passing rate (%) averaged over all beams per plan, shown for two different types of induced beam modeling errors. Gamma criteria was 1%/1 mm.

Moderate correlation/statistically significant cases having a positive slope, indicating that many of the larger critical errors in patient dose are occurring even when QA Gamma passing rates are high. In fact, some of the largest anatomy-based dose differences occurred in cases where the IMRT QA passing rates were 95%-100% (3%/3 mm Gamma). Instances of high IMRT QA passing rates despite high anatomy-based dose differences can be called false negatives, i.e., the IMRT QA Gamma passing rates, if taken alone, would lead one to conclude that there are no problems. Similarly, the results also show instances of low IMRT QA passing rates without any large differences in the anatomy-based dose metrics, which could be called false positives.

It can be concluded that Gamma passing rates, a very common ‘measured vs calculated’
QA performance metric, though perhaps useful in general commissioning of a system (TPS/delivery) or in catching gross errors, is clearly not sensitive to clinically relevant patient dose errors on a per-patient/plan basis.[29] This is consistent with the results of several other recent studies.[30–32] These findings call into question the value of conventional, per-beam QA methods employed for per-patient IMRT dose QA. First of all, it is intuitive that with per-patient dose errors, the importance is the location and overlap of these per-beam errors in terms of critical volumes (targets and organs at risk) and not about per-beam passing rates in a phantom. Take, as an example of a false negative, a hypothetical IMRT plan where there are small regions of ‘hot’ dose error in each field but not so large in size that the IMRT QA passing rate falls below, say, 95%. If those regions of higher-than-planned dose all overlap exactly at one portion of the spinal cord, there could be dire consequences. The anatomy-specific impact of these types of errors is not captured by conventional per-beam metrics, which do not have weight factors of errors with respect to anatomy intersections. As an example of a false positive, consider another hypothetical IMRT plan where each field exhibits low IMRT QA passing rates due to both hot and cold regions. However, suppose that these hot and cold regions do not overlap in any particular pattern in 3D and, rather, somewhat cancel each other (or at least dilute each other in magnitude) resulting in critical volume dose metrics that are not compromised beyond tolerance. In fact, in the case of a false positive, it is possible that the critical dose metrics of the true 3D patient dose may be superior to the planned dose. Figure 2.11 illustrates the regions of false positives (low QA passing rates but with noncritical patient dose errors) and regions of false negatives.
(high QA passing rates despite critical patient dose errors) which may be useful when examining Figs. 2.4-2.9.

**Figure 2.11:** Generalized illustration of regions of false negatives (high passing rates despite critical patient dose errors) and false positives (low QA passing rates but with noncritical patient dose errors) when correlating critical patient dose errors to conventional IMRT QA Gamma passing rates. In this schematic, the critical dose error threshold is “E” and the standard acceptance criteria for Gamma passing rates is “C.”

The conventional wisdom of percent difference/DTA-based criteria is based on commissioning TPS dose calculation algorithms and not on per-patient QA. In fact, when commissioning a treatment planning or dose delivery system, these conventional methods may be useful, as they provide quantified metrics that a physicist can use to optimize aspects of a beam model (or beam delivery) by comparing calculations to measurements (in phantom) and rigorously tuning the system for highest accuracy and consistency.
However, in per-patient IMRT dose QA, the DTA (and the even more lenient Gamma) might hide significant errors. The current standard of 3 mm DTA is quite large considering that today’s margins in the area of image guidance are often near this level. It is easily shown, for instance, that if one shifts a very conformal 3D dose grid by 3 mm, the planned DVHs become quite unacceptable even though the ‘3 mm’ criterion will still be met. Likewise, the use of global percent difference (normalize error percentages to the max dose in a plan or in a field) can hide significant low dose errors that may overlap in critical structures where an organ tolerance is already near its limit. The potential for such errors is illustrated by the error ranges in Table 2.2, as all of these observed errors happened for plans where the average IMRT QA passing rate was 95% or greater (3%/3 mm and 2%/2 mm). Employing 1%/1 mm criteria closed the tolerance substantially.

The observed potential for dose errors in critical anatomy and the clear lack of correlation are troubling. It is possible that an expected trend of correlation might appear if we had induced much larger or even catastrophic errors (wrong fields delivered, wrong beam energies, gross MU errors, etc.). Additional studies on larger induced errors and/or on different types of errors would be interesting. However, for the sake of this study, we focused exclusively on the QA passing rates that are above, or at least near, commonly accepted levels. It could be argued that any proposed QA standards (i.e., methods, performance metrics, acceptance criteria, tolerances, etc.) should be proven effective/sensitive/predictive before they are recommended as standards and their weaknesses and potential failures should be carefully documented. It is not valuable to keep searching for instances where the current methods ‘might work’ if enough evidence has been shown
to the contrary. The results of this work certainly call into question the utility of the 3%/3 mm Gamma passing rates as an adequate metric for per-patient IMRT QA.

This study has focused on per-field planar dose analysis, a method which is less relevant for 3D composite QA, for which 3D dosimetry phantoms are used and composite dose (from all sub-beams) is measured and compared to the TPS calculation of the plan-on-phantom. It is of vital importance that correlation studies should also be performed for these types of commercial phantoms with their varying detector locations. Intuitively, one might expect that the QA phantom dose analysis might only correlate with anatomy-based dose differences if the detectors overlap in 3D where the critical structures are, which is different for each patient. To the knowledge of the author, nothing has been proven yet either supporting or questioning the usefulness of these 3D phantom methods. In Chapter 3 a similar correlation is conducted to answer this question.

In conclusion, the study presented in this chapter shows there is a lack of correlation between conventional IMRT QA performance metrics (Gamma passing rates) and dose differences in critical anatomic regions-of-interest. The most common acceptance criteria and published actions levels therefore have insufficient, or at least unproven, predictive power for per-patient IMRT QA. Moreover, the methodology of basing action levels on prior performance achievements using these conventional methods is unwarranted because meeting these criteria does not ensure clinically acceptable dose errors.
Chapter 3

Correlation Study of 3D Gamma vs Patient DVH Error

3.1 Introduction

In Chapter 2 we have demonstrated that Gamma passing rate, when used in planar IMRT QA, lacks sensitivity and specificity in predicting clinically important patient dose errors. This result calls into question the utility of Gamma passing rate as a commonly adopted metric for the purpose of patient-specific dose QA.

However, although per-beam, planar dose verification is the most common way to perform Gamma-based IMRT QA[23], the use of advanced 3D techniques has been rapidly
growing in both research and clinical practice. Several newly developed radiation dosimeters are made with detector elements embedded in 3D geometries, namely the ‘ring geometry’, the ‘spiral geometry’, and ‘bi-planar geometry’. A brief introduction of the currently available 3D detectors is provided as follows:

- Delta4 (ScandiDos AB, 3 Uppsala, Sweden) uses a bi-planar detector geometry[33–36]. The detector consists of a cylindrical PMMA phantom with two orthogonal detector boards. A total of 1069 p-type 1 mm wide diode detectors are located in the measurement area of 20×20 cm² in each of the measurement planes.

- ArcCHECK (Sun Nuclear Corporation, Melbourne, FL)[37–39] consists of two cylindrical shells made of PMMA, which interlock to form a hollow cylinder with 15 and 25 cm inner and external diameters, respectively, and a length of 13.2 cm. A circuit board with 124 diodes is wrapped around the inner cylindrical shell forming four adjacent rings of detectors with 19.7 cm diameter, spaced 1 cm apart.

- The spiral film phantom invented by the Tomé lab in University of Wisconsin implements a spiral phantom design[40, 41]. A solid water cylindrical phantom of 30 cm diameter and 15 cm height was machined to create an archimedean spiral cavity for placing either radiographic or radiochromic film. The phantom was also machined in several locations to place an ionization chamber to verify the delivered dose at these points.

These new detector designs allow physicists to perform dose points verification that are sampled three dimensionally in the phantom. Several groups have reported their
clinical experience using these 3D detectors for IMRT QA[33, 34, 36, 41], and volumetric modulated arc therapy (VMAT) QA[37–39]. When comparing the planned and measured dose in the 3D phantoms, 3D Gamma passing rate (a natural generalization of the planar Gamma passing rate, also the form of the Gamma concept in the original paper), as well as its mathematically-modified version was used as the metric[14, 42–44].

Despite the fact that 3D Gamma analysis with these new detector designs is already being implemented for patient-specific dose QA, whether 3D Gamma has sufficient predictive power to detect patient dose error remains unproven. The previous chapter has shown that this is not the case in 2D, and that significant patient DVH errors could remain undetected if QA action levels are only based on planar Gamma passing rate. 3D Gamma might have an advantage over 2D Gamma since it samples more information spatially. However, it is still of great importance to perform a similar correlation study to answer this question.

From a purely scientific point of view, the previous chapter has focused on 2D, per-beam scenario, i.e. Gamma passing rates generated by comparison of the planned and measured phantom planar dose distribution for each single beam. To interpret the lack of correlation to patient DVH error in this scenario, one can immediately list the following three factors that could play a potential role: (1) 2D dose distribution actually undersamples the real data, as opposed to 3D, (2) a homogenous, regularly-shaped phantom is geometrically different from real patient, and (3), the intrinsic difference between the two metrics (Gamma passing rate and patient DVH error) caused the lack of correlation.
It is also of great interest to conduct further study to explore if and how any of these factors indeed degraded the correlation.

In this chapter, we performed a through investigation on the correlation between 3D Gamma passing rate and patient DVH error, with the consideration of both phantom Gamma and 3D patient (whole volume and per ROI) Gamma, in our effort to prove/disprove the utility of 3D Gamma for patient-specific dose QA, and to (partially) explore the question raised in the previous paragraph.

### 3.2 Materials and Methods

An experimental design similar to the one performed in the Chapter 2 was employed in this study. A brief summary is provided below, and also schematized in Fig 3.1. Twenty four clinically approved and treated head and neck IMRT patient data were chosen from our database. All patients were planned utilizing the Pinnacle TPS (Philips Radiation Oncology Systems, Fitchburg, WI) using 6 MV photon beams with 120-leaf MLC. Head and Neck IMRT plans are usually highly modulated with multiple irregular shaped region of interests (ROIs) and complex dose distribution, making them suitable for the purpose of this study.

In the TPS, an error-free beam model (clinically commissioned) was intentionally altered in four unique ways to generate the following four error-induced beam models: a high MLC transmission beam model, a low MLC transmission beam model, a shallow MLC penumbra beam model, and a very shallow MLC penumbra beam model. The detailed
Figure 3.1: Schematic of the experimental design. An error-free beam model was used to create ‘virtual measurements’ (to phantoms, to patient). An error-induced beam model was used to emulate an imperfect dose calculation, emulated to both phantom and patient. Gamma passing rates were produced for both 3D phantom (3 geometries) and patient dose.

A description of these beam models can be found in section 2.2. Again the degrees of these induced errors were tuned to result in phantom planar Gamma passing rates comparable to what is commonly achieved in the clinic. The magnitudes and types of resulting errors, including sample DVH curves can also be found in the previous chapter.
3.2.1 Virtual QA using 3D dosimeters with three distinct detector geometries

In the virtual QA scheme, the error-induced beam models serve as an ‘imperfect TPS’, and the error-free beam model represents the real machine delivery. For each patient, four IMRT plans with corresponding phantom QA plans were generated using each of the four error-induced beam models, resulting 96 IMRT patient plans. For each patient plan, a phantom plan was generated with the same error-induced beam model to obtain the planned phantom dose. Finally, all the patient plans and phantom plans were recalculated using the error-free beam model to simulate real delivery/measurement (serving as the ‘virtual measurement’ in this experiment). This technique has the extra advantage of removing the influence of prevalent measurement uncertainties of actual devices (2D arrays, 3D arrays, etc.), thus controlling the variables of the experiment and allowing us to study the metrics in their purest sense.

Phantom Gamma passing rates were generated by comparing the planned and ‘virtually measured’ phantom dose, with different criteria (3%/3 mm, 2%/2 mm, 1%/1 mm, both local and global percentage difference, lower threshold of analysis at 10% of global max).

Three different detector designs were evaluated. As illustrated in Fig 3.2, the phantom dose was calculated on a homogenous cylindrical phantom image, and to simulate each detector geometry, the ‘virtual measurements’ were filtered by three ROI’s corresponding to carefully designed cross-sections: namely a spiral, an ‘X’, and an ‘O’, mimicking commercial 3D dosimeters that have been described in section 3.1. Virtual measurement
phantom doses were calculated at 1 mm×1 mm×1 mm dose grid resolution, which is quite fine, but necessary in order to probe the more stringent Gamma criteria (such as 1%/1 mm). Patient doses were calculated at 4 mm×4 mm×4 mm. Our sensitivity test shows the 4 mm patient dose grids produced DVH-based metrics almost identical to 2 mm dose grids.
Figure 3.2: Sagittal, axial, and coronal dose planes for a representative dataset out of the 96 datasets studied. The error induced here was the very shallow MLC penumbra. (A) Planned (error-induced) patient dose; (B) true (error-free) patient dose; (C) difference map overlaid on planned dose; (D) calculated (error-induced) phantom dose; (E) virtual measurement (error-free), and (F) difference map overlaid on the calculated dose. In panels C and F, the dose differences are local percent different per voxel, calculated as: $100\% \times (\text{simulated measurement} - \text{calculated})/\text{calculated}$. 
3.2.2 Virtual QA using 3D patient dose grids

The virtual QA scheme allows us to calculate the ‘delivered’ full density patient dose in the original patient CT image, which is not feasible to obtain from a true measurement based experiment. From the planned and virtually delivered patient dose, we calculated, for each of the 96 QA cases, the full density 3D patient Gamma passing rate, both in the whole patient volume and filtered by each specific ROI. Three different sets of Gamma criterion ((3%/3 mm, 2%/2 mm, 1%/1 mm) were chosen for the Gamma calculation. Local percentage difference was used to increase the sensitivity.

To evaluate the clinical impact of each ‘virtual QA’ case, the change in six clinically important patient DVH endpoints (CTV D95, spinal cord D1cc and D_{max}, D_{mean} to each parotid, and Larynx D_{mean}) were generated by comparing the planed and ‘virtually measured’ patient DVH. Patient doses were calculated at 4 mm×4 mm×4 mm. Our sensitivity test shows the 4 mm patient dose grids produced DVH-based metrics almost identical to 2 mm dose grids.

To investigate the utility of 3D Gamma, a correlation analysis that is similar to the previous chapter was conducted for the DVH errors against each set of 3D Gamma passing rates (phantom, patient, ROI).
3.3 Result

3.3.1 Correlation of 3D phantom Gamma passing rates (three geometries) to DVH-based metrics

Figs 3.3-3.8 and Table 3.1 summarize the results of the 3D dosimeters correlation of Gamma passing rates (three criteria permutations: 3%/3 mm, 2%/2 mm, and 1%/1 mm for both local percent difference and global percent difference methods) vs each of the DVH-based metrics. Figs 3.3-3.8 plot only the 3%/3 mm Gamma and using the global normalization percent difference method, as these are the most commonly used criteria in the clinics[23], and also suggested by TG119[12]. The figures shows a considerable number of outliers, with some points falling in regions that could be considered false negatives.

As for the correlation analysis (Table 3.1), there were some interesting results:

- First, for the most common criteria, 3%(global)/3 mm, no detector geometry exhibited a Pearson correlation coefficient (r) that had even a moderate correlation (|r| > 0.4) and with the expected slope (r < 0) that would indicate that DVH errors decrease with increasing Gamma passing rate.

- Second, the same 3%(global)/3 mm Gamma criteria produced an abundance of points that might be considered ‘false negatives’ as evidenced by the phantom Gamma passing rates exceeding 95% in all cases but where DVH errors exceeded 5%.
**Figure 3.3:** Magnitude of errors in CTV D95 dose vs 3D detector Gamma passing rate, shown for three different detector designs (Gamma calculated at 3%/3mm criteria)

**Figure 3.4:** Magnitude of errors in spinal cord D1cc dose vs 3D detector Gamma passing rate, shown for three different detector designs (Gamma calculated at 3%/3mm criteria)
**Figure 3.5:** Magnitude of errors in spinal cord max dose vs 3D detector Gamma passing rate, shown for three different detector designs (Gamma calculated at 3%/3mm criteria)

**Figure 3.6:** Magnitude of errors in contralateral parotid mean dose vs 3D detector Gamma passing rate, shown for three different detector designs (Gamma calculated at 3%/3mm criteria)
Figure 3.7: Magnitude of errors in ipsilateral parotid mean dose vs 3D detector Gamma passing rate, shown for three different detector designs (Gamma calculated at 3%/3mm criteria)

Figure 3.8: Magnitude of errors in larynx mean dose vs 3D detector Gamma passing rate, shown for three different detector designs (Gamma calculated at 3%/3mm criteria)
• Third, in many cases and especially for the ‘X’ and Spiral detector geometries, the correlation did improve (i.e., |r| increased, with r<0) as Gamma criteria grew more stringent (3%/3 mm to 1%/1 mm), though in most cases the correlation was still weak (|r| >0.4).

• Fourth, as more stringent criteria were examined, the correlations for the ‘X’ geometry began to trend better than the spiral and even more so than the ‘O’ geometry.

• And fifth, there was a significant frequency of inverse correlations (in these cases, positive r-values) where the DVH errors were larger for higher Gamma pass rates (17 of 36 cases for ‘O’; 11/36 for ‘X’; 11/36 for spiral).

Note that the terms ‘false positive’ and ‘false negative’ discussed in this study are only considered in the sense of patient DVH errors. It is important to emphasize that any significant deviation of the original planned dose/DVH and a QA-derived dose/DVH deserves investigation, even if the QA metrics are more favourable than the original plan. ‘False positives’ in this study may indicate no adverse effect on a particular patient plan, but that does not mean the source of the underlying difference should be ignored. Any good manufacturing process will try to drive down all sources of variation, as what may be an advantageous error in one case may be a problem in another.
Table 3.1: Pearson product-moment correlation coefficient (r) and two-tailed p-values correlating the magnitude of anatomy dose errors to 3D detector Gamma passing rate for different detector geometries, calculated with three sets of Gamma criteria. We show results for both global percent and local percent difference methods, and the lower threshold for analysis was 10% of the global maximum dose.

<table>
<thead>
<tr>
<th></th>
<th>CTV D95</th>
<th>Spinal Cord D1cc</th>
<th>Spinal Cord Dmax</th>
<th>Contralateral Parotid Dmean</th>
<th>Ipsilateral Parotid Dmean</th>
<th>Larynx Dmean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>r</td>
<td>p-value</td>
<td>r</td>
<td>p-value</td>
<td>r</td>
<td>p-value</td>
</tr>
<tr>
<td>&quot;O&quot; Detector Gamma Metric</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3%/3mm</td>
<td>0.478</td>
<td>&lt;0.01</td>
<td>-0.386</td>
<td>&lt;0.01</td>
<td>0.113</td>
<td>0.27</td>
</tr>
<tr>
<td>2%/2mm</td>
<td>0.321</td>
<td>&lt;0.01</td>
<td>-0.358</td>
<td>&lt;0.01</td>
<td>0.052</td>
<td>0.62</td>
</tr>
<tr>
<td>1%/1mm</td>
<td>-0.274</td>
<td>&lt;0.01</td>
<td>-0.384</td>
<td>&lt;0.01</td>
<td>-0.243</td>
<td>0.02</td>
</tr>
<tr>
<td>Spiral Detector Gamma Metric</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3%/3mm</td>
<td>0.319</td>
<td>&lt;0.01</td>
<td>-0.459</td>
<td>&lt;0.01</td>
<td>0.039</td>
<td>0.71</td>
</tr>
<tr>
<td>2%/2mm</td>
<td>-0.113</td>
<td>0.30</td>
<td>-0.449</td>
<td>&lt;0.01</td>
<td>-0.128</td>
<td>0.21</td>
</tr>
<tr>
<td>1%/1mm</td>
<td>-0.804</td>
<td>&lt;0.01</td>
<td>-0.382</td>
<td>&lt;0.01</td>
<td>-0.532</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>&quot;X&quot; Detector Gamma Metric</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3%/3mm</td>
<td>0.361</td>
<td>&lt;0.01</td>
<td>-0.371</td>
<td>&lt;0.01</td>
<td>0.060</td>
<td>0.56</td>
</tr>
<tr>
<td>2%/2mm</td>
<td>-0.135</td>
<td>0.21</td>
<td>-0.442</td>
<td>&lt;0.01</td>
<td>-0.128</td>
<td>0.21</td>
</tr>
<tr>
<td>1%/1mm</td>
<td>-0.867</td>
<td>&lt;0.01</td>
<td>-0.382</td>
<td>&lt;0.01</td>
<td>-0.558</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>
3.3.2 Correlation of 3D patient Gamma passing rates (whole volume, per ROI) to DVH-based metrics

Fig 3.9-3.14 plot the errors in DVH-based metrics (the CTV D95, D1cc and Dmax for the spinal cord, and mean dose for contralateral parotid, ipsilateral parotid, and larynx) vs 3D Gamma passing rates. Two separate panels were plotted for each patient anatomy DVH under consideration, one using the Gamma passing rate of the whole patient dose grid, the other using the Gamma passing rate of only the corresponding ROI. Gamma indices were calculated using three different sets of parameters (3%/3 mm, 2%/2 mm, 1%/1 mm, with local percentage dose difference and 10% lower threshold cutoff). Each voxel passes if its Gamma index is calculated to be smaller than 1. It is seen that the 3%/3 mm Gamma passing rate of the whole patient dose grid has at best a weak correlation ($|r| < 0.4$ and $r$ with negative sign) with patient DVH deviations.

It is also observed that moving to ROI specific Gamma passing rate and utilizing more stringent criteria (2%/2 mm and 1%/1 mm) improves the correlation between Gamma passing rate and patient DVH deviations in most cases, as expected. However almost all correlations remain weak to moderate, as evidenced from the Pearson correlation coefficients listed in Table 3.2. The only exception is that at 1%/1mm, CTV D95 has a strong desired correlation with 3D Gamma passing rate, however 1% already exceeds the measurement uncertainty of common detectors, rendering this correlation clinically irrelevant.
Figure 3.9: Magnitude of errors in CTV D95 dose vs (a) 3D Gamma passing rate in the whole patient volume, and (b) 3D Gamma passing rate in CTV, shown for three different set of Gamma parameters (3%/3mm, 2%/2mm, 1%/1mm, local percentage difference)
Figure 3.10: Magnitude of errors in spinal cord D1cc dose vs (a) 3D Gamma passing rate in the whole patient volume, and (b) 3D Gamma passing rate in spinal cord, shown for three different set of Gamma parameters (3%/3mm, 2%/2mm, 1%/1mm, local percentage difference)
Figure 3.11: Magnitude of errors in spinal cord Dmax dose vs (a) 3D Gamma passing rate in the whole patient volume, and (b) 3D Gamma passing rate in spinal cord, shown for three different set of Gamma parameters (3%/3mm, 2%/2mm, 1%/1mm, local percentage difference)
Figure 3.12: Magnitude of errors in contralateral parotid mean dose vs (a) 3D Gamma passing rate in the whole patient volume, and (b) 3D Gamma passing rate in contralateral parotid, shown for three different set of Gamma parameters (3%/3mm, 2%/2mm, 1%/1mm, local percentage difference)
Figure 3.13: Magnitude of errors in Ipsilateral parotid mean dose vs (a) 3D Gamma passing rate in the whole patient volume, and (b) 3D Gamma passing rate in Ipsilateral parotid, shown for three different set of Gamma parameters (3%/3mm, 2%/2mm, 1%/1mm, local percentage difference)
Figure 3.14: Magnitude of errors in larynx mean dose vs (a) 3D Gamma passing rate in the whole patient volume, and (b) 3D Gamma passing rate in larynx, shown for three different set of Gamma parameters (3%/3mm, 2%/2mm, 1%/1mm, local percentage difference)
Table 3.2: Pearson product-moment correlation coefficient (r) and two-tailed p-values correlating the magnitude of anatomy dose errors to Gamma passing rates calculated by comparing 3D patient dose grids (i.e., in the hypothetical case of the patient as the ‘detector’), over three sets of Gamma criteria. In all cases, the percent difference normalization was local to maximize sensitivity, and the lower threshold for analysis was 10% of the global maximum dose.

<table>
<thead>
<tr>
<th>Whole patient Gamma Metric</th>
<th>CTV D95</th>
<th>Spinal Cord D1cc</th>
<th>Spinal Cord Dmax</th>
<th>Contralateral Parotid Dmean</th>
<th>Ipsilateral Parotid Dmean</th>
<th>Larynx Dmean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>r</td>
<td>p-value</td>
<td>r</td>
<td>p-value</td>
<td>r</td>
<td>p-value</td>
</tr>
<tr>
<td>3%/3 mm</td>
<td>0.179</td>
<td>0.10</td>
<td>-0.344</td>
<td>&lt;0.01</td>
<td>-0.343</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>2%/2 mm</td>
<td>-0.267</td>
<td>0.01</td>
<td>-0.381</td>
<td>&lt;0.01</td>
<td>-0.379</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>1%/1 mm</td>
<td>-0.793</td>
<td>&lt;0.01</td>
<td>-0.320</td>
<td>&lt;0.01</td>
<td>-0.308</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>3%/3 mm</td>
<td>0.107</td>
<td>0.32</td>
<td>-0.487</td>
<td>&lt;0.01</td>
<td>-0.360</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>2%/2 mm</td>
<td>-0.335</td>
<td>&lt;0.01</td>
<td>-0.498</td>
<td>&lt;0.01</td>
<td>-0.370</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>1%/1 mm</td>
<td>-0.86</td>
<td>&lt;0.01</td>
<td>-0.513</td>
<td>&lt;0.01</td>
<td>-0.387</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>
3.4 Discussion and Conclusion

3.4.1 Correlation of 3D phantom Gamma passing rates (three geometries) to DVH-based metrics

In this modeling study, we have found that conventional Gamma passing rates for three virtual 3D dosimeters with arrays of detectors arranged in common 3D geometries are not consistently correlated to clinically relevant patient DVH-based metrics. This agrees with the finding for per-beam planar pass rates from the previous chapter. It is interesting that the correlations do, in fact, improve as one sets more stringent criteria (i.e., changing from 3%/3 mm to 1%/1 mm); however, the pass rates at the 1%/1 mm level are very low and have a large range. Actions levels (for passing rates) presuppose a strong correlation, and such correlation was not found in this work. To further study the potential of the Gamma passing rate, we investigated the most favorable conditions for the Gamma index, i.e., the actual patient dose grids, both whole and filtered by ROI (these results are discussed in the following section).

It is worth mentioning that our study is focused specifically on the Gamma passing rate metric and does not imply that other uses of the Gamma index may not be useful. Certainly when commissioning a TPS or delivery system (when there are no patient ROIs or DVHs to consider), the Gamma passing rate is a metric of value. Moreover, if Gamma criteria are made more stringent (say, 2%/2 mm or less), the passing rate is a potentially useful metric to optimize for phantom dose analysis. Furthermore, our experiment represents ideal measurements of high-density detectors in the named geometries, and with
‘perfect’ measurements not subject to the measurement uncertainties and inaccuracies represented by actual 3D dosimeters. This allowed us to study the QA metrics in their purest sense by controlling those variables. Also designed out of this experiment where any and all dosimeter-specific variation in analyses methods, as we used a single analysis tool for all comparisons.

3.4.2 Correlation of 3D patient Gamma passing rates (whole volume, per ROI) to DVH-based metrics

Our results show that the 3D patient Gamma passing rate for the whole dose grid, although calculated in vivo and at full density, has weak correlation to errors in the DVH-based metrics. This lack of correlation in patient-specific IMRT QA can be understood if one thinks of the Gamma passing rate as a non-patient-specific metric. From a clinical point of view, all voxels in patient image do not have equal importance for radiotherapy treatment. The same dose errors can have big clinical impact if they coincide geometrically with critical structures, say, target volumes or OARs, as opposed to if they occur outside of critical structures. The shapes, sizes, and locations of critical volumes are patient-specific and variable. The physicist cannot determine the potential clinical impact without, essentially using, the same DVH-based metrics employed to approve the plan in the first place. Gamma passing rate for the whole dose volume, even for the actual patient anatomy, does not provide information about the anatomical location of were the failure occurs or at which dose level it failed, both of which are important.
To further investigate the limitation of Gamma passing rate in IMRT QA, information about anatomical location was incorporated into the Gamma passing rate in the form of a local Gamma passing rate, calculated specifically for each ROI. It is seen from the results that in most cases, the correlation to errors in DVH-based metrics was improved when using the corresponding ROI specific Gamma passing rate, as expected. However, almost all correlations were still either weak or moderate (with the exception of the CTV at 1%/1 mm). This could be explained by the fact that the Gamma passing rate, which although provides the quantity of errors, does not specify the magnitude of the error. For instance, if a 95% Gamma passing rate is reported for a serial organ, say, the spinal cord, what is immediately important is not whether 95% is high enough, but rather the magnitude and direction of the error for those 5% of failing voxels and their impact on the clinical relevant dose metrics, such as Dmax and D1cc, which cannot be told from the passing rate itself. The 5% failing voxels may or may not result in a change in Dmax and D1cc. Moreover the Gamma metric treats the percentage dose error in an absolute fashion while the sign of error is also clinically important. Continuing with the example of spinal cord, we observed in our study that in many cases, a low spinal cord Gamma passing rate could correspond to a decrease in Dmax and D1cc, which is actually advantageous in terms of spinal cord complication probability. These examples clearly indicate that lacking the information of the magnitude (and sign) of the errors is a hindrance to the Gamma passing rate and as a matter of fact, to passing rates in general.

Note that the lack of correlation between 3D Gamma passing rate and clinical relevant
metrics observed in this study allows us to address the questions raised in the introduction: namely that the lack of predictive power of planar Gamma passing rate analysis is not only due to the limitation of 2D sampling (as opposed to 3D), but also due to the intrinsic limitation of the Gamma passing rate itself, i.e., the lack of information on the location and magnitude (and sign) of dose errors.

### 3.4.3 Notes on Local Percent Difference and 2D Composite QA

When calculating 3D Gamma passing rate, almost every clinic use global percent difference and normalize the percentage dose difference to a global maximum or prescription[23]. However, it is shown from our results that the ‘global’ option might hide some relatively big errors for ROIs as these errors may not exceed the 3% threshold when normalized to the global maximum. With this in mind, the local percentage dose difference was examined (in addition to global percentage difference) in this study in an effort to catch errors in lower dose regions. Fig 3.15 (A) and (B) show the Gamma index distribution (3%/3 mm) using the two types of percentage difference on the same slice for a given patient. In this case, the ‘virtually delivered’ dose to the spinal cord region is around 1.3 Gy higher than the planned. It is seen that these errors are caught in the Gamma index map with local percentage difference, while hidden in the Gamma index map with global percentage difference. This is mainly because in this case, the dose difference is less than 3% of the normalization dose (75 Gy). In fact, the planned spinal cord $D_{max}$ in this particular case is 45.2 Gy and the actual spinal cord $D_{max}$ is 46.5 Gy. Although this difference does not exceed 3% of the normalization dose, it may still be significant. Also
of note is that when result of local percent difference was used, more ‘false positive’ cases were seen since some not-clinically important dose errors in the very low dose region are detected.

Another clinical relevant observation is that the planar Gamma index distribution sampled from the 3D patient Gamma index map varies significantly from slice to slice. As illustrated in Fig 3.15 (C) and (D), in one coronal slice the there are no failing voxels in the CTV, and Gamma looks flawed for the contralateral parotid, however in another coronal slice that is 1.5 cm away from the first slice towards the anterior direction, the situation is opposite: Gamma distribution looks flawed in CTV but perfect in the contralateral parotid. This observation agrees with a published study[45] that shows in 2D composite QA, “verification results are strongly dependent on the plane chosen”, and that “no correlation could be found between the levels of error in different verification planes”. The inconsistency of 2D composite Gamma distribution between different planes shows that it is not a robust method/metric.

In conclusion, Gamma passing rates for dose-in-phantom for common 3D dosimeter detector geometries exhibit negligible to weak correlation with errors in clinically relevant patient DVH metrics. Gamma passing rates based on in vivo dose grids, in 3D full density for the whole patient volume and filtered by ROI, exhibit negligible to weak correlation at the 3%/3 mm and 2%/2 mm levels, with instances of moderate correlation at 1%/1 mm. The Gamma passing rate, due to the lack of correlation to patient DVH error, is not a sufficient metric for patient-specific dose QA.
Figure 3.15: Gamma index failing voxel distributions (colour) overlaid on the TPS dose (grayscale). ((A) and (B)) Case 15, LTBM error, 3%/3 mm on a coronal slice, calculated with: (A) global percentage difference and (B) local percentage difference. (C) and (D)) Case 04, LTBM error, 3%/3 mm, local percentage difference in two coronal slices separated by 1.5 cm: (C) contours of CTV (the larger contour on the right) and (D) contralateral parotid (the smaller contour on the top left) are displayed.
Chapter 4

Evaluation of a Patient Dose Prediction Method

4.1 Introduction

The purpose of patient-specific dose QA is to ensure the quality of the delivered radiation treatment matches to what is planned, in terms of tumor coverage and normal tissue sparing. Therefore, the ideal way of performing patient-specific dose QA would be to measure the dose in the actual patient, and compare the measured with the planned dose. However, this is obviously not feasible due to ethical issues, and from the technical perspective it is not possible to measure dose in the whole patient volume with a high density. For these reasons any measurement-based patient-specific QA has to be performed with a phantom dosimeter. However, dose measured in a phantom usually
does not have any direct clinical implications (as compared to dose in the patient). As a result, planned and measured phantom doses are usually compared in the general sense of ‘how well the two distributions match’, not ‘how well the clinical impact of the two dose distributions match’. Gamma passing rate is a mathematically simple metric that is ideal for comparison in the former sense, hence has been widely adapted in the clinics. However, the previous two chapters have demonstrated that the Gamma passing rate is not predictive of clinical impact, thus may not be a sufficient metric for the purpose of patient-specific dose QA.

In this chapter and the next one, we aim to explore new QA methods/metrics. As been discussed in previous chapters, any good patient-specific QA metric should be based on information about the patient dose. Although we mentioned that it is not directly measurable, there are several existing commercially available systems and research approaches that claim the ability to predict the delivered patient dose from phantom measurement/QA records. A brief introduction of these approaches is provided below:

- The Compass system (IBA-Wellhofer) allows incident fluence-based patient dose reconstruction[46–48]. An ion chamber array (Matrixx, IBA-Wellhofer) is mounted on the gantry facing the beam direction during the QA delivery and the detector response at each time segment is recorded. These detector response is then converted to the incident radiation fluence at each gantry angle. With the measured fluence, the patient dose is recalculated using an independent collapsed-cone algorithm on the patient CT.
• 3DVH (Sun Nuclear) also detects incident fluence using a diode array detector (MapCheck, Sun Nuclear), but instead of recalculating the patient dose, it utilizes a perturbation algorithm that perturbs the planned patient dose to obtain an estimation of the real delivered patient dose\cite{49, 50}. A detailed description of this algorithm will be provided later in this chapter.

• Dosimetry Check (Math Resolution LLC, Columbia, MD) performs patient dose reconstruction with transit fluence\cite{51, 52}. It deconvolves the response of the Electric Portal Imaging Device (EPID) during the real patient treatment to acquire the transit fluence at each gantry angle. A back-project calculation is then carried onto the patient CT to reconstruct the patient dose.

• MLC log file based patient dose reconstruction is an ongoing research approach that uses the log file of the record and verify (R&V) system to recalculate the patient dose\cite{53}. MLC log file records detailed MLC motion information during the treatment\cite{54, 55}. This ‘real’ MLC information is used to determine the actual beam fluence, and in turn recalculate the dose on the patient CT.

These patient dose prediction methods, once verified to be accurate and reliable, can revolutionize the method/metric of patient-specific QA. Instead of comparing the dose in the phantom in terms of Gamma passing rates, physicist and physicians will be able to directly compare the planned and estimated patient dose, in terms of potential clinical impact on the patients. Therefore, it is of great interest to evaluate the capacity of these patient dose prediction methods.
In this chapter, we conduct a ‘virtual QA’ study to verify the patient dose prediction accuracy of one the approaches listed above (3DVH, Sun Nuclear).

4.2 Materials and Methods

4.2.1 Virtual QA Scheme with 3DVH

A virtual QA scheme that is similar to the previous two chapters was conducted. In the TPS, an error-free beam model (serves as the ‘virtual machine’) was intentionally altered in four different ways (MLC transmission high and low, MLC penumbra shallow and very shallow), resulting in four error-induced beam models. These four error induced models serve as ‘imperfect TPS ’ in this virtual QA scheme. For each of the 24 clinically approved H&N patients, four IMRT plans were created using each of the error induced models. These plans were then recalculated with the error-free beam for ‘virtual delivery’. Then the per-beam planar phantom dose was calculated using both the error-induced beam model (planned planar QA dose) and the error-free beam model (virtually measured planar QA dose).

The 3DVH software was used to predict the patient dose from QA measurements. This software make use of an algorithm called Planned Dose Perturbation (PDP), that takes into account the planned and measured per-beam planar QA dose, as well as the planned patient dose, to produce an estimation of the delivered patient dose. A detailed description of this method is provided in the next section.
Finally, the PDP estimated patient dose/DVH was compared to the virtually delivered patient dose/DVH, in order to evaluate the performance of this patient dose prediction method.

4.2.2 Patient dose/DVH prediction with PDP

The major function of the PDP algorithm is to use conventional per-beam planar dose QA methods to feed a sophisticated three-dimensional (3D) perturbation system that ‘corrects’ the original 3D patient dose as generated by the TPS and outputs a 3D patient dose grid that has built into it the manifestation of any errors detected by the planar QA. While some systems use a third party dose engine and estimated fluence (derived from EPID images or measurement arrays)[51], PDP uses perturbation methodology designed specifically for Compton effects of high energy photons. The PDP method does not introduce new sources of variation or error that may occur with an independent 3D dose algorithm (i.e., variations that might not be errors but just differences of the new algorithm vs the TPS algorithm). PDP will alter dose only if and where dose differences are detected in conventional dosimetry array systems. The perturbation methodology is summarized here for completeness, as provided by and with permission of the vendor.

Per-beam 2D error masks

As with conventional planar QA, the planned and ‘virtually delivered’ phantom planar doses are paired for each treatment beam in the plan at the exact treatment MU.
However, conventional Gamma analyses and passing rates are of no consequence to PDP. Rather, the absolute dose planes are superimposed (and optionally autoshifted to account for minor setup errors) to produce two-dimensional (2D) ‘error masks’ where absolute dose differences (measured - calculated) and local percentage errors are stored. These are created in batch and saved for subsequent consumption by the perturbation engine. The local percentage difference (measured-calculated QA dose) error mask for the i-th beam is called

\[ \Delta \text{QA}_{\text{dose}}(i, j)_{\text{QA}_{\text{depth, distance}}} \]  

**Partitioning of per-voxel contribution factors**

The DICOM RT plan, structure set, and planned dose (3D) are exported from the TPS, which are then imported into the PDP software. The treatment beams in the RT plan are associated with their corresponding planar QA dose pair (and thus with the error mask). From the beam geometries relative to the patient volume and from the beams’ control points and MU settings, PDP derives the per-voxel dose contribution from each beam. Thus, for the \( i \)-th beam the contribution at dose voxel \((x, y, z)\) is estimated using a physics-based energy deposition model. Call this specific voxel’s dose contribution from the \( i \)-th beam: \( \text{Dose}_{xyz}^i \). Dose voxel contributions are partitioned per-beam. A built-in PDP beam model (specific to the treatment machine, energy, and MLC-type) is used for this step and also for the next (rendering of errors masks into the patient volume).
Rendering of the error masks into patient volume

Each dose voxel in the 3D dose grid, having been partitioned into per-beam contributions, can be ‘perturbed’ using the PDP algorithm. The error mask is modified (depending on depth in patient and distance of the voxel from the source) by a contribution modifying function (CMF) and applied to the dose contribution of the beam in question. The CMF: (1) computes the contribution of voxel dose from each beam so that individual per-beam patient dose or control point doses are not necessary; and (2) modifies the error from the specific phantom geometry and radiological depth to that of the patient. The CMF is based on prestored models (rendered for each linac vendor, model, and energy) using, among other parameters, TERMA parameters and Monte Carlo dose kernels. For the \(i\)-th beam and a single voxel \((xyz)\) this dose perturbation can be described as

\[
PDP_{xyz}^i = \Delta QAdose(i, j)^i_{QAd, depth, distance} \\
\times CMF(depth_{xyz}^i, distance_{xyz}^i) \\
\times QAdose(i, j)^i_{QAd, depth, distance} \times Dose_{xyz}^i
\]  

Accumulating per-voxel errors over all voxels, all beams

Accumulating the total dose perturbation over all voxels and beams can then be used to modify (perturb) the input patient dose grid calculated by the TPS, which includes
all necessary heterogeneity dependencies. The total perturbation is therefore

\[ PDP_{\text{plan}}^{\text{volume}} = \sum_{xyz} \sum_{i} PDP_{i}^{xyz} \quad (4.3) \]

and this perturbation can be applied to the TPS dose grid to generate the predicted patient dose grid that is corrected for known errors measured in the QA phantom geometry. To evaluate PDP for the purpose of DVH based IMRT QA, patient DVH estimations were then generated from the predicted patient dose and compared to the ‘actual’ patient DVH values generated from the ‘virtually delivered’ patient dose.

### 4.3 Results

Figs 4.1-4.6 show the scatter plots of actual patient DVH error (difference between error-free and error-induced) vs the PDP predicted patient DVH error (difference between PDP prediction and error-induced), for each of the six DVH based metrics studied in the previous chapters. Linear regression analysis was also performed for each DVH metric, with the slope, the y-intercept, and the coefficient of determination \( R^2 \) reported in corresponding figures. It is seen that the predicted DVH metrics are as a whole very close to the actual value as all the linear regression slopes are near one (range: 0.907 to 1.042), the y-intercepts are all near zero, and the values for \( R^2 \) are all greater than 0.94 (an ideal fit would give a slope of one, and intercept of zero, and a value of \( R^2 = 1 \)). Table 4.1 lists the correlation coefficients and corresponding p-values between the actual and PDP predicted patient DVH errors *in the last row*. 
**Figure 4.1:** Actual vs PDP-predicted patient DVH error for CTV D95, with the results of linear regression fitting.

**Figure 4.2:** Actual vs PDP-predicted patient DVH error for spinal cord D1cc, with the results of linear regression fitting.
Figure 4.3: Actual vs PDP-predicted patient DVH error for spinal cord Dmax, with the results of linear regression fitting.

Figure 4.4: Actual vs PDP-predicted patient DVH error for contralateral parotid mean dose, with the results of linear regression fitting.
Figure 4.5: Actual vs PDP-predicted patient DVH error for ipsilateral parotid mean dose, with the results of linear regression fitting.

Figure 4.6: Actual vs PDP-predicted patient DVH error for larynx mean dose, with the results of linear regression fitting.
Table 4.1: Pearson product-moment correlation coefficient (r) and two-tailed p-values correlating the PDP predicted and actual patient DVH errors in the last row. The correlation coefficients between 3D patient Gamma and patient DVH error from the previous chapter is provided (with gray background) for comparison.

<table>
<thead>
<tr>
<th></th>
<th>CTV D95</th>
<th>Spinal Cord D1cc</th>
<th>Spinal Cord Dmax</th>
<th>Contralateral Parotid Dmean</th>
<th>Ipsilateral Parotid Dmean</th>
<th>Larynx Dmean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>r</td>
<td>p-value</td>
<td>r</td>
<td>p-value</td>
<td>r</td>
<td>p-value</td>
</tr>
<tr>
<td>&quot;O&quot; Detector Gamma Metric</td>
<td>3%/3mm</td>
<td>0.478</td>
<td>&lt;0.01</td>
<td>-0.386</td>
<td>&lt;0.01</td>
<td>0.113</td>
</tr>
<tr>
<td></td>
<td>2%/2mm</td>
<td>0.321</td>
<td>&lt;0.01</td>
<td>-0.358</td>
<td>&lt;0.01</td>
<td>0.052</td>
</tr>
<tr>
<td></td>
<td>1%/1mm</td>
<td>-0.274</td>
<td>&lt;0.01</td>
<td>-0.384</td>
<td>&lt;0.01</td>
<td>-0.243</td>
</tr>
<tr>
<td>Spiral Detector Gamma Metric</td>
<td>3%/3mm</td>
<td>0.319</td>
<td>&lt;0.01</td>
<td>-0.459</td>
<td>&lt;0.01</td>
<td>0.039</td>
</tr>
<tr>
<td></td>
<td>2%/2mm</td>
<td>-0.113</td>
<td>0.30</td>
<td>-0.449</td>
<td>&lt;0.01</td>
<td>-0.128</td>
</tr>
<tr>
<td></td>
<td>1%/1mm</td>
<td>-0.804</td>
<td>&lt;0.01</td>
<td>-0.382</td>
<td>&lt;0.01</td>
<td>-0.532</td>
</tr>
<tr>
<td>&quot;X&quot; Detector Gamma Metric</td>
<td>3%/3mm</td>
<td>0.361</td>
<td>&lt;0.01</td>
<td>-0.371</td>
<td>&lt;0.01</td>
<td>0.060</td>
</tr>
<tr>
<td></td>
<td>2%/2mm</td>
<td>-0.135</td>
<td>0.21</td>
<td>-0.442</td>
<td>&lt;0.01</td>
<td>-0.128</td>
</tr>
<tr>
<td></td>
<td>1%/1mm</td>
<td>-0.867</td>
<td>&lt;0.01</td>
<td>-0.382</td>
<td>&lt;0.01</td>
<td>-0.558</td>
</tr>
<tr>
<td>PDP Prediction</td>
<td>0.993</td>
<td>&lt;0.01</td>
<td>0.983</td>
<td>&lt;0.01</td>
<td>0.969</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>
Fig 4.7 shows a comparison of the distributions of the differences between the PDP estimated and the true DVH-based metrics for each of the DVH-based metrics we have considered using box plots, which allow one to compare these distributions in a convenient manner in terms of their center, spread, skew, and the length of their tails. All of them have a median value that is close to 0 Gy and a tight interquartile range (≤0.3 Gy). However, for both the skew and the length of tails of the distributions no general pattern can be discerned. Half of the distributions are left skewed, while the other half are right skewed with the minimum and maximum values extending less than 0.75 Gy from the median.

**Figure 4.7**: Box plots of the difference between the PDP-estimated DVH-based metrics and the actual DVH-based metrics of the error-free standards. The horizontal line in middle shows the median (50th percentile) of the distribution. The top and bottom of the box show the 75% and 25% percentile, respectively. While the top and bottom whiskers extend to the minimum and maximum values, respectively.
Figure 4.8 shows, for one representative patient and three induced error-types, the distribution of dose differences on a dose point-by-point basis for all dose points >10 Gy (≈49,000 points analyzed per plan). The ratio of (error-induced/error-free) quantifies the error magnitude per dose point, and the ratio of (PDP/error-free) quantifies how accurate the PDP correction predicts the true (error-free) patient dose. It is seen that from the per point dose error perspective, MLC transmission modeling error causes on average ±4% point dose error, and PDP was able to predict it to on average within ±0.1%. PDP also shrinks the standard deviation of error distribution by half in the MLC penumbra error case.

4.4 Discussion and Conclusion

The PDP predicted patient DVH has been shown to be in good agreement with the actual values, indicating that it is possible to use the sampled phantom measurement data to produce metrics that correlate well with DVH differences. These results demonstrate the feasibility of utilizing direct prediction of the patient DVH instead of Gamma passing rate analysis for patient-specific IMRT QA. In fact, the essence of patient-specific IMRT QA is to ensure that the dose distribution that is going to be delivered to the patient is of the same/comparable quality as the approved plan, and such quality is evaluated by patient dose statistics and DVH curves. The evolution from Gamma passing rates to DVH-based metrics thus natural.
Figure 4.8: Dose error distributions before correction (error-induced vs error-free) and after correction (PDP vs error-free) with errors represented by a ratio of absolute dose per point. All dose points 10 Gy and above were analyzed. Examples shown for one representative patient, with: (A) High transmission beam modeling error; (B) Low transmission error; and (C) Very shallow penumbra error.
During the first decade of clinical IMRT (prior to practical and commercial techniques of direct patient dose prediction from QA measurement) Gamma passing rates have served IMRT QA as an imperfect but simple metric, and acceptance criteria have commonly been under the jurisdiction of the medical physicist. With the development of new IMRT QA devices and software focused on patient dose estimation, a simple passing rate is no longer applicable. Although some have investigated the use of 3D Gamma analysis based on the estimated patient dose[52], it is shown in Chapter 3 that these patient Gamma passing rates are not sufficient for patient-specific dose QA. Pass/fail decisions must now be based, generally speaking, on the difference between the planned patient dose and DVH and the QA-system-estimated patient dose and DVH, and the potential clinical impact.

Subsequently, in lockstep with accurate DVH-estimation capabilities comes the need for evolution in the design of the QA processes, i.e., the approval mechanisms and action levels for QA metrics will need to be revisited. On one hand, this is at least an intuitive concept because QA metrics can be based on the same metrics used for plan approval and overall plan quality assessment. Generally speaking, a clinic could set action levels on critical DVH goals, and if those goals were met by the original treatment plan but are not met by the estimated QA DVHs, then it would require mitigation prior to QA approval. On the other hand, these critical DVH goals could grow to be numerous and complex (compared to a simple passing rate metric), and this could introduce inefficiencies in the busy clinical environment. In the extreme case, a physician might be required to review the patient dose and DVHs twice -once upon completion of the TPS plan and again after
the pre-treatment dose QA. This could be impractical at some institutions. Therefore, instead of the raw DVH value, other simpler and clinically relevant metrics that based on the patient dose and DVH prediction should be visited for their potential use in pre-treatment dose QA. In chapter 5 we will investigate the usage of radiobiological treatment outcome models for this purpose.

In conclusion, the patient dose prediction QA method we investigated in this study was shown to accurately predict the delivered patient DVH using conventional planar QA results. Therefore it is possible to move from Gamma passing rate based to patient dose prediction based pre-treatment dose QA. Using patient-DVH-based metrics IMRT QA will allows per-patient dose QA to be sensitive and specific to clinically relevant patient dose errors. Further studies are required to analyze specific new QA methods/metrics based on patient dose prediction for their effectiveness and practicality in the clinical setting.
Chapter 5

The Use of Radiobiological Models in Patient-Specific QA

5.1 Introduction

In Chapter 4 we validated the patient dose/DVH prediction accuracy of a QA algorithm/software (PDP/3DVH) and obtained promising results. The success of these new methods potentially allows clinical QA metrics to be built on the predicted patient dose, instead of the measured phantom dose. Some studies reported the use of using 3D in vivo Gamma analysis for patient dose verification[52] . However, although it might be a useful tool to commission a QA system, Chapter 3 suggests that in patient-specific QA, 3D Gamma passing rates are not correlated to clinically important patient DVH
changes, rendering it an insufficient QA metric. New QA metrics that can more effectively represent clinical impact need to be proposed.

Radiobiological models could serve as one potential base for such new metrics. These models are commonly-used tools that describe the relationship between the delivered radiation dose and the expected clinical outcomes, such as the Tumor Control Probability (TCP), Normal Tissue Complication Probability (NTCP), and Uncomplicated Tumor Control Probability (UTCP). Both TCP models and NTCP models usually take a sigmoid shape in describing the radiation response, as demonstrated in Fig 5.1[56]. The UTCP, or probability of complication free cure, can then be derived from the TCP and NTCP models.

![Sigmoid-shaped curves for tumor control and normal tissue complication probabilities, adapted from Holthsen 1936](image)

**Figure 5.1:** Sigmoid-shaped curves for tumor control and normal tissue complication probabilities, adapted from [Holthsen 1936]

There are two general approaches to create a TCP model: theoretically based (mechanistic) and empirically based (phenomenological). Theoretically based models are built
on the existing knowledge about cancer stem cell survival after radiation, which can be approximately characterized by the Linear Quadratic (LQ) equation\[57\]. One can then apply Poisson statistics to determine the probability of local tumor control, which is essentially the occurrence of zero cancer stem cell survival. From this basic model, several modifications can be made to incorporate the effect of inhomogenous tumor stem cell distribution\[58\], inhomogenous dose distribution\[58\], and tumor proliferation\[59, 60\]. The advantage of these mechanistic models is that they are easier to interpret, and more compatible to additional factors.

In contrast, phenomenological TCP models assume no knowledge of radiobiology, but are derived from fitting curves to actual clinical data. Because of this, the parameters in these models need to be determined specifically for each individual disease. These models usually have simpler mathematical forms compared to mechanistic models, and are more flexible in terms of parameter-tuning. One widely used phenomenological TCP model is the logit model\[61\];

\[
P(D)_{\text{Logit}} = \left[ 1 + \left( \frac{D_{50}}{D} \right)^{4\gamma_{50}} \right]^{-1}
\]  

(5.1)

where \(D_{50}\) is the dose which yields a tumor control probability of 50%, while \(\gamma_{50}\) is the normalized slope (ratio of TCP change and percent dose change) at \(D_{50}\).

Although there are also mechanistic approaches for NTCP modeling, the most commonly used NTCP model is a phenomenological model developed by Lyman\[62\]. A significant consideration in NTCP modeling is that unlike tumor targets, usually only part of the
normal tissue volume is irradiated during the treatment. This so called “volume effect”
was taken into account in the Lyman model by introducing the volume effect parameter
\( n \)[61]. Kutcher and Burman then extended the original Lyman model to incorporate
heterogeneous dose distribution[63], resulting in the well known LKB model;

\[
NTCP(gEUD) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \exp\left(-\frac{u^2}{2}\right)du
\]

\[
t = \frac{gEUD - D_{50}(1)}{mD_{50}(1)}; \quad gEUD = \left(\sum_{i} v_i D_i^{\frac{1}{n}}\right)^n
\]

where \( v_i \) is the normalized volume that has a dose level of \( D_i \), \( TD_{50} \) is the dose, if
irradiated uniformly on the entire organ, would cause a 50\% chance of normal tissue
complication. Here \( m \) is the slope parameter, and \( n \) is the volume effect parameter.
The inhomogeneity of the dose distribution is taken into account using the generalized
equivalent dose \((gEUD)\) concept[64].

These radiobiological models, when applied to the planned and predicted patient dose,
will give a change in TCP and NTCP (\( \Delta TCP \) and \( \Delta NTCP \)), which may serve as a new
metric to identify clinically relevant dose errors from the QA results. In this chapter,
we will conduct a virtual QA study to explore the utility of \( \Delta TCP \) and \( \Delta NTCP \) in
patient-specific dose QA.
5.2 Materials and Methods

5.2.1 Error simulation

To simulate the difference between treatment planning and delivery, ten different types of errors were induced into a gold standard beam model in the Pinnacle (Philips Radiation Oncology Systems, Fitchburg, WI) treatment planning system. These errors can be categorized into either a TPS modeling error or a machine delivery error. 4 of the 10 induced errors were described in the previous chapters, however for the completeness of the content, a brief description of all ten induced errors is provided as follows:

TPS Modeling Error

- MLC transmission error

  The MLC transmission factor in the gold standard beam model is altered to simulate error in MLC transmission modeling. Two error-induced beam models are generated using this method: a high-transmission beam model that has an MLC transmission factor (3.88%), which is doubled of the error-free value (1.94%), and a low-transmission beam model that has an MLC transmission factor (0.97%), which is half of the error-free value.

- MLC penumbra error

  The field edge penumbra parameters in the gold standard beam model are altered to simulate error in MLC penumbra modeling. Two error-induced beam models are generated using this method: a shallow-penumbra beam model that has an
MLC penumbra (80%-20%) of 7.2 mm (at $d_{max}$) and 9.2 mm (at 10 cm depth), and a very-shallow-penumbra beam model that has an MLC penumbra (80%-20%) of 8.6 mm (at $d_{max}$) and 11 mm (at 10 cm depth).

- MLC tongue and groove (T&G) width error The T&G width in the gold standard beam model is altered to simulate error in MLC T&G modeling. Two error-induced beam models are generated using this method: a wide-T&G beam model that has an MLC T&G width of 2 mm, which is wider than that of the error free beam model (1 mm), and a thin-T&G beam model that has an MLC T&G width of 0.05 mm. As suggested by Pinnacle Physics reference guide, 2 mm and 0.05 mm are the upper and lower limits for MLC T&G width modeling.

**Machine Delivery Error**

- Output

The output factor in the gold standard beam model is altered to simulate error in machine output. Two error-induced machines are generated using this method: a high-output machine that has an output per MU 3% higher than the gold standard machine, and a low-output machine that has an output per MU 3% lower than the gold standard machine.

- MLC position error (systematic shift)

Pinnacle scripts were used to induce systematic MLC position error in the patient plan in two ways: opening all the open leaves in each MLC bank by 1mm (resulting a 2mm increase in the gap between each pair of MLC), and closing all the open
leaves in each MLC bank by 1mm (resulting in a 2mm decrease in the gap between each pair of MLC.

5.2.2 Virtual QA scheme

This study examines twenty head and neck (H&N) patients and twenty prostrate patients that were clinically approved and treated with step-and-shoot IMRT. For each patient, ten ‘virtual QA’ cases were created to simulate the effect of the ten different errors. For TPS modeling error simulation, the error-induced beam model was used to generate the patient plan. This plan was then recalculated using the gold-standard model to simulate delivery. For machine delivery error simulation, the gold-standard beam was used to generate the patient plan. This plan was then recalculated with the induced errors to simulate delivery. As a result, for each of the forty patients, ten sets of plan and QA doses were generated.

5.2.3 Data analysis

\( \Delta \text{TCP} \) and \( \Delta \text{NTCP} \) for different error types

The DVHs of eight different Regions of Interest (ROI) (CTV, GTV, contralateral parotid, ipsilateral parotid, spinal cord, and larynx for H&N cases, CTV and rectal wall for prostate cases) were exported directly from TPS using scripts for each set of patient doses. TCP for targets and NTCP for several Organs at Risk (OAR) were then calculated based on the DVHs. For TCP calculation, we chose a two parameter phenomenological
TCP model that is based on the logit model introduced in section 5.1, and modified by Kim and Tomé\cite{65};

\[
TCP\{\{D_i, v_i\}\} = \prod_{i=1}^{N} \left[ \frac{1}{1 + \left(\frac{D_{50}}{D_i}\right)^{\gamma_{50}}} \right]^{v_i}, \text{ where } \prod_{i=1}^{N} v_i = 1
\] (5.3)

This voxel based TCP model assumes a homogenous tumor (uniform radio-sensitivity throughout all tumor voxels) with an inhomogeneous dose distribution, which is mathematically convenient and fits the purpose of this study well. In this expression, the input \(v_i\) is the normalized volume that has a dose level of \(D_i\). \(D_{50}\) is (again) the dose which yields a tumor control probability of 50%, while \(\gamma_{50}\) is the normalized slope (ratio of TCP change and percent dose change) at \(D_{50}\). The values of these parameters in this study were chosen from literature. For H&N cases, \(D_{50}=50.44\text{ Gy, } \gamma_{50}=1.83\) for CTV and \(D_{50}=63.43\text{ Gy, } \gamma_{50}=2.66\) for GTV\cite{66}. While for prostate CTV, \(D_{50}=70.5\text{ Gy, } \gamma_{50}=2.9\)\cite{67}. For hypofractionation prostate cases, the patient dose was adjusted for the fraction size (assuming \(\alpha/\beta=3\text{ Gy}\)) to 2 Gy fractions.

The widely accepted three parameter Lyman-Kutcher-Burman (LKB) model, as introduced in Section 5.1 was chosen for NTCP calculation for the spinal cord, both parotids, larynx, and rectal wall. The parameter values were chosen as follows: \(TD_{50}=35\text{ Gy, } n=1, m=0.46\) for parotid\cite{68}, \(TD_{50}=66.5\text{ Gy, } n=0.05, m=0.175\) for spinal cord\cite{69}, \(TD_{50}=46.3\text{ Gy, } n=0.45, m=0.16\) for larynx\cite{70}, and \(TD_{50}=76.9\text{ Gy, } n=0.09, m=0.13\) for rectum\cite{71}. 

A TCP/NTCP calculation module with graphic user interface (GUI) was created using MATLAB (The Mathworks Inc., Natick, U.S.A.). The common TCP/NTCP models as well as a data base for model parameters were programed in this module. Figure 5.2 demonstrates the user interface of this module. For each set of planned and QA patient dose, TCP and NTCP were calculated for corresponding ROIs. The change in TCP and NTCP for each virtual QA case was then calculated and analyzed for each ROI and each type of induced error.

**Correlation between $\Delta$DVH and $\Delta$TCP&$\Delta$NTCP**

Several commonly used DVH endpoints, namely CTV D95 (for both H&N and prostate cases), GTV $D_{95}$ (H&N cases only), spinal cord D1cc, larynx mean dose, and rectal wall V75 and V65, were chosen to be evaluated in this study. For each set of patient doses, corresponding DVH endpoint values were calculated for both the ‘planed dose’ and the ‘virtually delivered’ dose. A statistical analysis was conducted to determine correlation between the change in the DVH endpoint values and the change in corresponding TCP&NTCP, using Pearson’s correlation coefficients. A linear regression was also conducted to determine the ratio of change in TCP/NTCP to change in DVH metrics for all significant correlations observed.
**Figure 5.2:** A demonstration of the Matlab module for TCP/NTCP calculation.

![TCP-NTCP calculation module](image)

<table>
<thead>
<tr>
<th>TCP Cal</th>
<th>Fx size</th>
<th>a/b</th>
<th>Model</th>
<th>norm fx</th>
<th>D50</th>
<th>r50</th>
<th>TCP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Import DVH</td>
<td>2</td>
<td>10</td>
<td>Logistic</td>
<td>2</td>
<td>63.43</td>
<td>2.86</td>
<td>0.84067</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NTCP Cal</th>
<th>OAR</th>
<th>a/b</th>
<th>Model</th>
<th>norm fx</th>
<th>TD50</th>
<th>m/r</th>
<th>n</th>
<th>NTCP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Import DVH</td>
<td>Larynx</td>
<td>3</td>
<td>LKB</td>
<td></td>
<td>35</td>
<td>0.48</td>
<td>1</td>
<td>0.24411</td>
</tr>
<tr>
<td>Import DVH</td>
<td>Parotid</td>
<td>3</td>
<td>LKB</td>
<td></td>
<td>46.3</td>
<td>0.18</td>
<td>0.45</td>
<td>0.53808</td>
</tr>
<tr>
<td>Import DVH</td>
<td></td>
<td></td>
<td>LKB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Import DVH</td>
<td></td>
<td></td>
<td>LKB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Import DVH</td>
<td></td>
<td></td>
<td>LKB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---Heming Zhen---U of Wisconsin-Madison---hmzhen@gmail.com---
5.3 Results

5.3.1 ΔTCP and ΔNTCP for different error types

Tables 5.1-5.2 show the change in TCP/NTCP for each of the eight ROIs considered in this study, categorized by the type of induced errors. The average change in TCP/NTCP, as well as the range and standard deviation are listed for each ROI & error type combination. To help interpret these tables, figs 5.3-5.10 are presented to show the DVH curves of each of the eight selected ROIs, for all the patient plans created using the error-free model. Figs 5.11-5.18 show the points on the TCP/NTCP curves that represent these DVH data. For example, Each DVH curve for H&N CTV from fig 5.3 will be represented by one data point (marked as ‘x’) on the TCP curve for H&N CTV, as shown in fig 5.11.

The DVH curves for CTV and GTV tend to be more condensed together when compared to OARs, since they are the main planning objective and the prescription levels are similar between patients. This is also represented by the generally more condensed distribution of the data points on the TCP curves than on the NTCP curves. The indications of this observation will be discussed in the following sections.

The results in Tables 5.1-5.2 can be summarized as the following two facts:

- Different types of induced errors have different impacts on TCP and NTCP, and the magnitudes of these impacts vary among ROIs. For example, it is seen in CTV, GTV, ipsilateral parotid, and rectal wall 3% machine output error causes
Figure 5.3: DVH curves of H&N CTV, for all H&N patients.

Figure 5.4: DVH curves of H&N GTV, for all H&N patients.
Figure 5.5: DVH curves of contralateral parotid, for all H&N patients.

Figure 5.6: DVH curves of ipsilateral parotid, for all H&N patients.
Figure 5.7: DVH curves of spinal cord, for all H&N patients.

Figure 5.8: DVH curves of larynx, for all H&N patients.
Figure 5.9: DVH curves of prostate CTV, for all prostate patients.

Figure 5.10: DVH curves of rectal wall, for all prostate patients.
Figure 5.11: Range of H&N CTV TCP, for all H&N patients.

Figure 5.12: Range of H&N GTV TCP, for all H&N patients.
Figure 5.13: Range of contralateral parotid NTCP, for all H&N patients.

Figure 5.14: Range of ipsilateral parotid NTCP, for all H&N patients.
Figure 5.15: Range of spinal cord NTCP, for all H&N patients.

Figure 5.16: Range of larynx NTCP, for all H&N patients.
more change in TCP/NTCP than does MLC transmission error, while the opposite is true in contralateral parotid and larynx.

- For the same type of induced error, different organs experience different impacts on their TCP/NTCP. An example is that the 3% machine output error seems to cause greater change in TCP/NTCP for prostate CTV, H&N GTV and the rectal wall, than for other OARs. On the other hand, MLC penumbra modeling error seems to cause more TCP/NTCP change for contralateral parotid, larynx, and rectal wall, than for other ROIs. All these observations will be revisited in the following sections with further discussion.
Figure 5.18: Range of rectal wall NTCP, for all prostate patients.
Table 5.1: Change in TCP (CTV, GTV) and NTCP (both parotids, spinal cord, and larynx) in H&N cases, summarized by types of induced errors. The average change in TCP/NTCP, as well as the standard deviation and range are listed.

<table>
<thead>
<tr>
<th></th>
<th>Output High*</th>
<th>Output Low*</th>
<th>MLC Gap Wide*</th>
<th>MLC Gap Narrow*</th>
<th>T&amp;G wide*</th>
<th>T&amp;G thin*</th>
<th>Pen S*</th>
<th>Pen SS*</th>
<th>Tran High*</th>
<th>Tran Low*</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTV</td>
<td>2.17% ± 0.48%</td>
<td>-2.63% ± 0.54%</td>
<td>1.22% ± 0.65%</td>
<td>-1.44% ± 0.84%</td>
<td>0.16% ± 0.19%</td>
<td>0.06% ± 0.19%</td>
<td>0.05% ± 0.35%</td>
<td>0.21% ± 0.10%</td>
<td>-1.30% ± 0.37%</td>
<td>1.30% ± 0.41%</td>
</tr>
<tr>
<td>[1.50%, 3.17%]</td>
<td>[-3.71%, -1.86%]</td>
<td>[0.42%, 2.94%]</td>
<td>[-3.48%, -0.47%]</td>
<td>[-0.27%, 0.46%]</td>
<td>[-0.10%, 0.60%]</td>
<td>[-1.42%, 0.24%]</td>
<td>[-0.02%, 0.40%]</td>
<td>[-1.84%, 0.47%]</td>
<td>[0.49%, 2.20%]</td>
<td></td>
</tr>
<tr>
<td>GTV</td>
<td>4.18% ± 0.92%</td>
<td>-5.23% ± 0.91%</td>
<td>1.99% ± 1.17%</td>
<td>-2.55% ± 1.84%</td>
<td>0.31% ± 0.32%</td>
<td>0.10% ± 0.34%</td>
<td>0.66% ± 0.65%</td>
<td>1.11% ± 0.41%</td>
<td>-2.17% ± 0.57%</td>
<td>2.17% ± 0.58%</td>
</tr>
<tr>
<td>[3.34%, 6.53%]</td>
<td>[-7.49%, -4.27%]</td>
<td>[0.74%, 5.15%]</td>
<td>[-7.67%, -0.81%]</td>
<td>[-0.18%, 1.00%]</td>
<td>[-0.36%, 0.98%]</td>
<td>[-1.67%, 1.47%]</td>
<td>[-0.39%, 1.92%]</td>
<td>[-3.28%, -1.2%]</td>
<td>[1.16%, 3.49%]</td>
<td></td>
</tr>
<tr>
<td>Contralateral Parotid</td>
<td>1.19% ± 0.60%</td>
<td>-1.16% ± 0.59%</td>
<td>2.08% ± 1.25%</td>
<td>-1.80% ± 0.95%</td>
<td>-0.16% ± 0.27%</td>
<td>0.32% ± 0.26%</td>
<td>-0.85% ± 0.52%</td>
<td>-1.34% ± 0.75%</td>
<td>-2.01% ± 0.88%</td>
<td>2.36% ± 1.00%</td>
</tr>
<tr>
<td></td>
<td>[0.18%, 2.47%]</td>
<td>[-2.46%, -0.17%]</td>
<td>[0.30%, 5.52%]</td>
<td>[-3.83%, -0.28%]</td>
<td>[0.56%, 0.56%]</td>
<td>[0.18%, 0.68%]</td>
<td>[-1.94%, -0.15%]</td>
<td>[-3.05%, -0.26%]</td>
<td>[-3.54%, -0.39%]</td>
<td>[0.49%, 4.09%]</td>
</tr>
<tr>
<td>Ipsilateral parotid</td>
<td>1.73% ± 0.83%</td>
<td>-1.92% ± 0.85%</td>
<td>1.48% ± 1.22%</td>
<td>-1.49% ± 0.99%</td>
<td>0.14% ± 0.12%</td>
<td>0.11% ± 0.15%</td>
<td>0.14% ± 0.34%</td>
<td>0.17% ± 0.55%</td>
<td>-1.39% ± 0.97%</td>
<td>1.50% ± 1.10%</td>
</tr>
<tr>
<td>[0.83%, 2.78%]</td>
<td>[-2.88%, -1.03%]</td>
<td>[0.49%, 5.19%]</td>
<td>[-4.00%, -0.57%]</td>
<td>[-0.06%, 0.40%]</td>
<td>[-0.06%, 0.49%]</td>
<td>[-0.43%, 0.82%]</td>
<td>[-1.70%, 1.94%]</td>
<td>[-3.27%, -0.50%]</td>
<td>[0.55%, 3.69%]</td>
<td></td>
</tr>
<tr>
<td>Cord</td>
<td>0.10% ± 0.08%</td>
<td>-0.07% ± 0.07%</td>
<td>0.08% ± 0.06%</td>
<td>-0.05% ± 0.04%</td>
<td>0.01% ± 0.02%</td>
<td>0.00% ± 0.01%</td>
<td>-0.07% ± 0.05%</td>
<td>-0.11% ± 0.08%</td>
<td>-0.07% ± 0.06%</td>
<td>0.11% ± 0.08%</td>
</tr>
<tr>
<td>[0.00%, 0.04%]</td>
<td>[-0.32%, 0.00%]</td>
<td>[0.00%, 0.28%]</td>
<td>[-0.18%, 0.00%]</td>
<td>[0.01%, 0.05%]</td>
<td>[0.03%, 0.02%]</td>
<td>[0.22%, 0.00%]</td>
<td>[-0.31%, 0.00%]</td>
<td>[-0.30%, -0.00%]</td>
<td>[0.00%, 0.40%]</td>
<td></td>
</tr>
<tr>
<td>Larynx</td>
<td>0.76% ± 1.18%</td>
<td>-0.61% ± 0.99%</td>
<td>1.42% ± 2.12%</td>
<td>-0.88% ± 1.37%</td>
<td>-0.02% ± 0.10%</td>
<td>0.16% ± 0.26%</td>
<td>-1.17% ± 1.78%</td>
<td>-1.27% ± 1.99%</td>
<td>-0.88% ± 1.49%</td>
<td>1.25% ± 1.94%</td>
</tr>
<tr>
<td>[0.00%, 4.20%]</td>
<td>[-3.59%, 0.00%]</td>
<td>[0.00%, 7.08%]</td>
<td>[-4.82%, 0.00%]</td>
<td>[-0.26%, 2.22%]</td>
<td>[-0.05%, 0.79%]</td>
<td>[-6.39%, -0.01%]</td>
<td>[-7.37%, -1.27%]</td>
<td>[-5.35%, 0.00%]</td>
<td>[0.00%, 7.05%]</td>
<td></td>
</tr>
</tbody>
</table>

*Output High: machine output 3% high
Output Low: machine output 3% low
MLC Gap Wide: MLC gap opens by 2 mm (1mm on each bank)
MLC Gap Narrow: MLC gap closes by 2mm (1mm on each bank)
T&G wide: TPS tongue and groove width too wide
T&G thin: TPS tongue and groove width too thin
Pen S: TPS MLC penumbral shallow
Pen SS: TPS MLC penumbral too shallow
Tran High: TPS MLC transmission too high
Tran Low: TPS MLC transmission too low
Table 5.2: Change in TCP (CTV) and NTCP (Rectal wall) in prostate cases, summarized by types of induced errors. The average change in TCP/NTCP, as well as the standard deviation and range are listed.

<table>
<thead>
<tr>
<th></th>
<th>Machine Delivery Error</th>
<th>TPS Beam Modeling Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Output High*</td>
<td>Output low*</td>
</tr>
<tr>
<td><strong>CTV</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.44% ± 0.49%</td>
<td>-5.70% ± 0.55%</td>
</tr>
<tr>
<td></td>
<td>[3.75%, 5.24%]</td>
<td>[6.57%, -4.90%]</td>
</tr>
<tr>
<td><strong>Rectal Wall</strong></td>
<td>5.02% ± 0.95%</td>
<td>-4.09% ± 0.87%</td>
</tr>
<tr>
<td></td>
<td>[2.76%, 6.62%]</td>
<td>[5.65%, -2.09%]</td>
</tr>
</tbody>
</table>

*Output High: machine output 3% high
Output Low: machine output 3% low
MLC Gap Wide: MLC gap opens by 2 mm (1mm on each bank)
MLC Gap Narrow: MLC gap closes by 2mm (1mm on each bank)
T&G wide: TPS tongue and groove width too wide
T&G thin: TPS tongue and groove width too thin
Pen S: TPS MLC penumbra shallow
Pen SS: TPS MLC penumbra too shallow
Tran High: TPS MLC transmission too high
Tran Low: TPS MLC transmission too low
5.3.2 Correlation between $\Delta$DVH and $\Delta$TCP&$\Delta$NTCP

Table 5.3 summarizes the correlation between the change in DVH metric for each ROI studied and the change in their corresponding TCP or NTCP value, as demonstrated by the Pearson’s correlation coefficients as well as their two-tailed p-values. Each correlation coefficient was generated with the combined data for all twenty patients and all ten types of induced error (200 data points). The linear regression coefficient $\beta$ that represents the ratio of change in TCP/NTCP to change in DVH metrics are also listed. Figures 5.19-5.21 demonstrate the scatter plots of a very strong correlation, a strong correlation, and a moderate correlation.

All the DVH metrics have strong to moderate correlation to the change in corresponding TCP and NTCP, with different linear correlation coefficients. Specifically, changes in CTV D95, GTV D95 and contralateral parotid mean dose seem to have very strong and significant correlation ($r>0.95$) to changes in TCP/NTCP for both H&N and prostate cases. Changes in ipsilateral parotid mean dose, rectal wall V75 and V65, and spinal cord D1cc have strong correlation ($r>0.70$) to the changes in corresponding NTCP. On the other hand, change in larynx mean dose has a moderate correlation to the change in its corresponding NTCP ($r<0.70$).
Table 5.3: Pearson correlation coefficient (r) with corresponding two-tailed p-values correlating change in TCP/NTCP and change, as well as linear regression coefficient \( \beta \), for all selected DVH metrics.

<table>
<thead>
<tr>
<th>DVH metric</th>
<th>Correlation to ( \Delta \text{TCP} &amp; \Delta \text{NTCP} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( r )</td>
</tr>
<tr>
<td>HN_CTV D95</td>
<td>0.959</td>
</tr>
<tr>
<td>HN_GTV D95</td>
<td>0.964</td>
</tr>
<tr>
<td>Contralateral parotid mean dose</td>
<td>0.977</td>
</tr>
<tr>
<td>Ipsilateral parotid mean dose</td>
<td>0.893</td>
</tr>
<tr>
<td>Spinal cord D1cc</td>
<td>0.793</td>
</tr>
<tr>
<td>Larynx mean dose</td>
<td>0.631</td>
</tr>
<tr>
<td>Prostate CTV D95</td>
<td>0.982</td>
</tr>
<tr>
<td>Rectal wall V75</td>
<td>0.926</td>
</tr>
<tr>
<td>Rectal wall V65</td>
<td>0.860</td>
</tr>
</tbody>
</table>

Figure 5.19: Scatter plot of change in prostate CTV TCP vs. change in prostate CTV D95, under all types of induced errors.
Figure 5.20: Scatter plot of change in ipsilateral parotid NTCP vs. change in ipsilateral parotid mean dose, under all types of induced errors.

Figure 5.21: Scatter plot of change in larynx NTCP vs. change in larynx mean dose, under all types of induced errors.
5.4 Discussion

5.4.1 The use of ΔTCP and ΔNTCP as patient-specific QA metric

An ideal QA metric should directly reflect the quality that needs assuring. In the realm of patient-specific QA, the quality that needs assuring is not a physical dose distribution, but the clinical impact of a plan on the patient. In other words, the purpose of patient-specific QA is not solely to demonstrate that the physical dose distribution does not deviate much from the planned dose, but more importantly, to demonstrate that the expected outcome of the treatment (based on the delivered dose) does not deviate much from the planned outcome. The conventional Gamma based metrics lacks clinical meaning: when a certain Gamma passing rate is obtained during a patient-specific dose QA, it is inherently difficult to draw any conclusion on how much the clinical outcome of the treatment could be affected. In fact, recent studies[29, 49] have shown that the same Gamma passing rates could correspond to different change in patient DVH values, and they may flag out errors that are not likely to have a big impact on the patient.

What’s more, Gamma passing rates could miss errors that will actually degrade the quality of the treatment. On the other hand, TCP and NTCP models are created to predict treatment outcomes. ΔTCP and ΔNTCP have direct clinical meanings – the change in the possibility of tumor control and normal tissue complication. These metrics can flag out errors with actual clinical impact, and therefore are more suitable for the purpose of patient-specific dose QA.
Recall fig 2.11 from Chapter 2, conventional Gamma based QA creates a lot of false negatives and false positives. The power of the new metric is to allow us remove these false negatives and false positives. On one hand, the new metric will help us catch clinically relevant patient dose errors that are missed by Gamma passing rate. On the other hand, it will let us pass the false positive cases. As a result, utilizing the new metric can save the time for physicists to concentrate on the errors that are of real clinical importance. It should also be made clear that if a big dose or DVH error is found, even if the new metrics show no clinical impact on TCP or NTCP, something needs to be investigated. If the problem is systematic, it will eventually impact a different patient plan.

5.4.2 $\Delta$TCP and $\Delta$NTCP for different error types

In this chapter we have intentionally induced error sources in IMRT treatment and summarized the impact on the expected clinical outcome – TCP and NTCP. It is helpful to think that these induced errors take effect on TCP/NTCP in two steps: the induced errors first cause changes in the delivered physical dose, and then the dose changes of each ROI will in turn cause changes in their corresponding TCP and NTCP.

In the first step, an error source causes physical dose changes in the ROIs, and depending on the interplay between the treatment geometry and the characteristic of the error source, the impact of a certain error source on the physical dose of different ROIs can vary. The first observation in section 5.3.1 partially proves this point. We observed that 3% machine output error causes more change in TCP/NTCP than MLC transmission
error for CTV, GTV, ipsilateral parotid and rectal wall, but less change in TCP/NTCP for contralateral parotid and larynx. Because TCP and NTCP are monotonically increasing functions, this actually indicates that 3% machine output error causes bigger change in dose (EUD/gEUD) for the former four ROIs than MLC transmission error, and less for the latter two ROIs. Since the output change will uniformly affect the whole treatment volume, and MLC transmission error will only affect the regions under the MLC, and since, during the treatment, the contralateral parotid and larynx are more likely to be under the MLC than the target and OAR that are nearer to the target (such as the ipsilateral parotid and rectal wall), they will be relatively more affected by MLC transmission errors.

In the second step, the change in physical dose leads to changes in TCP/NTCP for the ROIs. Depending on the local slope of the TCP/NTCP curve (which is determined by the shape of the curve and the dose level of the original plan), the same level of dose change may cause different level of TCP/NTCP change. The 3% output error serves as a good example: a 3% output error will theoretically increase the dose everywhere by 3%, and thus will induce the same level of EUD/gEUD change for every ROI. However, the impact on TCP/NTCP for different ROIs varies significantly. Specifically, they are higher for prostate CTV, H&N GTV and rectal wall than for other ROIs. If we go back and examine Figs 5.11-5.18, it’s not hard to tell that for these ROIs the original plans all fall into a region with relatively steep slopes, while for other ROIs the original plans may spread into regions with shallower slopes. Fig 5.16 demonstrates that the cord dose in virtually all plans lies in the very flat heel region of the cord NTCP curve, which is
why neither type of induced error causes significant change in spinal cord NTCP.

5.4.3 Indication on pre-treatment assessment of plan robustness

The discussion in the previous section naturally leads to another interesting and important topic: evaluating the robustness of the plan before treatment. It is important to note that both TCP and NTCP curves follow a sigmoid shape. As evidenced from the results above, plans that fall into different regions of the curve have different qualities in terms of both goodness and robustness. Figs 5.22-5.23 provide a general demonstration of the different regions on TCP and NTCP curve. For TCP, the ideal plan should fall in the green region, since the TCP is high and most robust to perturbation, as evidenced by the flatness of the slope. The yellow region is less optimal, as the curve becomes steeper and plan is more sensitive to perturbation. The red region is not ideal because the TCP is either too low to begin with or too fragile to perturbation (steep slope). Similar logic can be used on NTCP curve. The green region represents plans that have low and stable NTCP, while the yellow region indicates less robustness, and the red region indicates the NTCP is either too high or too sensitive too dose perturbation, as demonstrated in Fig 5.23. These general demonstrations provide the basis of using the concept of TCP and NTCP to build tools to assess the robustness of a plan. Such tools can be used to examine how concrete a plan is under any type of error that perturbs the patient dose; they are not limited to the errors that are induced in this study. Here the word ‘robustness’ is used in the sense of the estimated clinical outcome, i.e. TCP and NTCP. Errors will inevitably happen in the radiotherapy treatment; knowing how robust a plan
is before treatment might help physicists and physicians decide whether/what actions should be taken. It would also be interesting to integrate the concept of plan robustness into the treatment planning process to improve the quality of radiotherapy plans in the first place.

\[ \text{Dose} \]

**Figure 5.22:** General demonstration of different regions on a TCP curve. Green region means TCP is high and most robust, yellow region indicates TCP is less optimal and more sensitive to errors, red regions means TCP is either too low or to fragile to perturbation.

5.4.4 Correlation between $\Delta$DVH and $\Delta$TCP&$\Delta$NTCP

It is reassuring to see that the change in all the selected DVH metrics correlates to the change in their corresponding TCP/NTCP, since these DVH endpoints are commonly used in the clinics to examine radiotherapy plans. However, unlike TCP/NTCP, DVH is still a physical concept that lacks clinical meaning. To interpret how much of a TCP/NTCP change will be caused by a change in DVH, one needs to know the ratio
between the change in TCP/NTCP and change in DVH, which is different for different
ROIs, as demonstrated in Table 5.3. In fact, these values are governed by the local slope
on the TCP/NTCP curve, and again will vary if the original plan falls in different region
of these curves. For example, if an original plan has a spinal cord D1cc of 10 Gy, then
a $+100\%$ change in this DVH metric might not cause any notable increase in NTCP;
however, if the cord D1cc is already pushing its tolerance, then even a $+2\%$ change in
D1cc might induce a significant change in NTCP. The scatter of data in fig 5.20 and
fig 5.21 also indicates the uncertainty of using a change in DVH to predict a change
in TCP/NTCP, even if their correlation is strong. Therefore, although correlated to
TCP/NTCP, DVH values are of limited use in terms of interpreting clinical impact, and
can not be a surrogate to TCP and NTCP.
To conclude this chapter, we have demonstrated, through a ‘virtual QA’ scheme with induced TPS modeling and machine delivery errors, the possibility of utilizing $\Delta$TCP and $\Delta$NTCP as patient-specific dose QA metrics. We have also demonstrated the potential use of TCP and NTCP models for evaluating the robustness of a treatment plan.
Chapter 6

Conclusion and Future Work

6.1 Is What We’re Doing Adequate?

This work for this dissertation has focused on patient-specific dose QA for modern radiation therapy treatment. Conventionally such QA is performed by comparing the planed and measured phantom dose. Clinical action levels have been based on the Gamma passing rates generated from this comparison. However, it has not yet been proven whether this metric is sensitive to patient dose errors. The first part of this work (Chapter 2-3) was performed to answer this important question: is the current patient-specific QA method/metric predictive of clinically important patient DVH errors? By inducing MLC transmission and MLC penumbra errors in the TPS, we have conducted 96 simulated QA cases. From the result of the correlation analysis, we found out that the planar Gamma passing rates does not correlate to errors in clinically important
patient DVH values, and that a lot of false negative and false positive cases occur. To investigate if this is just the deficiency of planar QA, since it only compares dose in one plane instead of the whole volume, in Chapter 3 this virtual QA study was repeated for 3D Gamma passing rates. The lack of correlation was found to be true for 3D composite Gamma as well, both in the phantom and in the whole patient volume (even filtered by ROI). Therefore we concluded that the Gamma passing rate itself is not strongly correlated to patient DVH errors, and may not, therefore, be a sufficient metric in catching clinically important patient dose errors. The conventional method/metric falls short for the purpose of patient-specific dose QA.

6.2 Is It Possible to Change?

Ideally patient-specific QA should be performed based on the patient dose instead of a phantom dose, and the action levels should be based on metrics that directly reflect the impacts on the predicted treatment outcome. The second part of this dissertation aimed to explore new method/metric in the vision of these criteria. In Chapter 4 we evaluated a commercially available QA system that is capable of predicting the patient dose from a phantom measurement, and have seen that the predicted patient dose/DVH is very accurate. The successful validation of this software as well as other patient dose prediction approaches (as discussed in Chapter 4) opens the possibility of new metrics that are based on patient dose/DVH. In Chapter 5 we explored this possibility by conducting virtual QA with more types of induced errors (MLC T&G width error, output
error, MLC position error), and using $\Delta$TCP and $\Delta$NTCP as the new metric to quantify the clinical impact of these errors. This new method/metric provides information that directly predicts the clinical impact on the patient, in terms of tumor control and normal tissue complication. We have also explored the potential of using TCP and NTCP models as tools to assess the robustness of the treatment plans. Moving towards these new QA methods/metrics will allow physicists and physicians to make more confident decisions on QA results, and better ensure the quality of treatment and safety of the patients. The new method/metric will allow us to reduce the false positives created by the conventional method/metric, and at the same time identify the false negatives and concentrate on errors that are likely to have a big clinical impact.

6.3 What’s Next?

The mission of providing better QA is a never ending quest for physicists. A natural follow-up of this work is the clinical implementation of patient-specific QA using patient dose prediction radiobiological metrics. Although the simulation work has shown promising results, much more need to be addressed before this methodology enters routine clinical practice. First is the commissioning of a patient dose prediction QA system. Chapter 4 has shown some TPS-based validation work; however, commissioning such a system through measurement can be tricky, since it is challenging to obtain the ground truth - the actual patient dose. Additional uncertainties in actual measurement (e.g., detector response variation, phantom positioning errors) also need to be controlled so
that the total uncertainty of the patient dose prediction process do not exceed the actual error that needs to be detected. Second, special attention is required in choosing the appropriate radiobiological models, as well as the model parameters for each specific organ and clinical endpoint. For this purpose a comprehensive database is required to store up-to-date TCP and NTCP models as well as their parameters. Chapter 5 has provided a foundation for this. It should also be realized that all the model parameters were obtained through fitting of clinical data, which only have a limited dose range, therefore, the model prediction of TCP/NTCP outside the dose range of the fitting data might be less reliable, especially for phenomenological models.

Chapter 5 also showed the possibility of using the concept of TCP and NTCP to evaluate plan robustness, which is a very interesting and important direction to explore in the future. Specifically, we will define the robustness of a treatment plan in a quantitative way. Such definition will combine the under-perturbation-characteristics of all clinically important ROIs, while these under-perturbation characteristics will be determined from the local change in TCP/NTCP under dose errors. This tool could be beneficial for treatment planning and pre-treatment plan evaluation purpose. In treatment planning systems, a plan robustness factor could be used as an objective or a penalty function in the optimization algorithm, to force the TPS to generate more robust plans. Physicists and physicians can use this tool to evaluate the robustness of a certain treatment plan before treatment, which will help them determine the risk of significant degradation of plan quality, and take better-informed actions.

A side product of Chapter 5 is the quantification of the clinical impacts of several different
types of TPS modeling and machine delivery errors. These impacts vary significantly between different types of errors. This result can lead to work in the broader realm of QA. The traditional paradigm of the machine QA programs is to check everything that can be checked. However, with the rapid development of radiotherapy machines and devices, the old paradigm is no longer feasible. It is time to bring in new ideas to answer the questions “What do we need to check?” and “When do we need to check?” Some efforts were made to adapt quality management (QM) concepts from the engineering field. QM tools like Failure Mode and Effect Analysis (FMEA)[72–75] and control charts[24, 76] have been investigated both by individual research groups and the on-going AAPM Task Group 100. One systematic way to answer the questions of “What to check” and “when to check,” is to quantify the actual risk of each error mode based on its severity, occurrence and detectability, as suggested by the FMEA method. The result in Chapter 5 is a summary of the ‘severities’ of several different error types. For future work, more sources of error can be included in the QA simulation and an FMEA for the mechanical process of MLC-based IMRT can be performed.
Bibliography


