Assessing the uncertainty in a normal tissue complication probability difference (ΔNTCP): radiation-induced liver disease (RILD) in liver tumour patients treated with proton vs X-ray therapy

Keiji Kobashi1,*,†, Anussara Prayongrat2,†, Takuya Kimoto3, Chie Toramatsu4, Yasuhiro Dekura2, Norio Katoh5, Shinichi Shimizu6,7, Yoichi M. Ito8 and Hiroki Shirato7,9

1Department of Medical Physics, Hokkaido University Hospital, North-15 West-7, Kita-ku, Sapporo, 0608638, Japan
2Department of Radiation Oncology, Graduate School of Medicine, Hokkaido University, North-15 West-7, Kita-ku, Sapporo, 0608638, Japan
3Department of Radiology, Kyoto Prefectural University of Medicine, 465 Kajii-cho, Kawaramachi-Hiroyoji, Kamigyo-ku, Kyoto, 6028566, Japan
4Department of Radiation Oncology, Tokyo Women’s Medical University, 8-1, Kawada-cho, Shinjuku, Tokyo, 1628666, Japan
5Department of Radiation Oncology, Hokkaido University Hospital, North-15 West-7, Kita-ku, Sapporo, 0608638, Japan
6Department of Radiation Oncology, Faculty of Medicine, Hokkaido University, North-15 West-7, Kita-ku, Sapporo, 0608638, Japan
7Global Station for Quantum Biomedical Science and Engineering, Global Institute for Cooperative Research and Education, Hokkaido University, North-15 West-7, Kita-ku, Sapporo, 0608638, Japan
8Department of Biostatistics, Faculty of Medicine, Hokkaido University, North-15 West-7, Kita-ku, Sapporo, 0608638, Japan
9Department of Radiation Medicine, Faculty of Medicine, Hokkaido University, North-15 West-7, Kita-ku, Sapporo, 0608638, Japan
*Corresponding author. Department of Medical Physics, Hokkaido University Hospital, North-15 West-7, Kita-ku, Sapporo, 0608638, Japan.
Tel: +81-11-706-5977; Fax: +81-11-706-7876; Email: kkobashi@med.hokudai.ac.jp
†These authors contributed equally to this work.

(Received 23 October 2017; revised 19 February 2018; editorial decision 20 February 2018)

ABSTRACT

Modern radiotherapy technologies such as proton beam therapy (PBT) permit dose escalation to the tumour and minimize unnecessary doses to normal tissues. To achieve appropriate patient selection for PBT, a normal tissue complication probability (NTCP) model can be applied to estimate the risk of treatment-related toxicity relative to X-ray therapy (XRT). A methodology for estimating the difference in NTCP (ΔNTCP), including its uncertainty as a function of dose to normal tissue, is described in this study using the Delta method, a statistical method for evaluating the variance of functions, considering the variance–covariance matrix. We used a virtual individual patient dataset of radiation-induced liver disease (RILD) in liver tumour patients who were treated with XRT as a study model. As an alternative option for individual patient data, dose-bin data, which consists of the number of patients who developed toxicity in each dose level/bin and the total number of patients in that dose level/bin, are useful for multi-institutional data sharing. It provides comparable accuracy with individual patient data when using the Delta method. With reliable NTCP models, the ΔNTCP with uncertainty might potentially guide the use of PBT; however, clinical validation and a cost-effectiveness study are needed to determine the appropriate ΔNTCP threshold.

Keywords: normal tissue complication probability; uncertainty; proton therapy; radiation-induced liver disease; liver tumour

© The Author(s) 2018. Published by Oxford University Press on behalf of The Japan Radiation Research Society and Japanese Society for Radiation Oncology. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
INTRODUCTION
Proton beam therapy (PBT) has emerged as a promising radiotherapy modality due to its favourable physical properties, in that it allows dose escalation to the tumour and minimizes unnecessary doses to normal tissues. With the aim of appropriate patient selection for PBT, a group of researchers at the University of Groningen, Netherlands, first introduced the model-based approach (MBA) to identify patients who could potentially benefit from PBT over X-ray therapy (XRT) by using a normal tissue complication probability (NTCP) model to estimate the risk of developing toxicity [1]. Langendijk et al. and Jokabi et al. investigated the NTCP difference (ΔNTCP) obtained from in silico planning, comparing intensity-modulated radiotherapy (IMRT) with intensity-modulated proton therapy (IMPT) in head and neck cancers [1, 2]. The selective use of PBT relies on a predefined ΔNTCP threshold, such as 10% for Grade II or 15% for total clinical benefit, according to the Dutch Society of Radiation Oncology (NRVO) guidelines [3]. However, these models were derived from a statistical assumption based on a small subset of the population, and the model uncertainty affects the accuracy of PBT selection [4]. In other words, underestimating ΔNTCP can eliminate the opportunity to benefit from PBT, whereas an overly cautious practice might cause the unnecessary use of this high-cost treatment. Therefore, uncertainty regarding the ΔNTCP values is necessary for better clinical implementation in the general population.

In the development of a reliable and generalized NTCP model, using large patient cohorts from multi-institutional datasets can further enhance the model accuracy. However, access to individual dose–volume histogram (DVH) data might be limited in local institutions due to ethical or technical issues. Recently, Wedenberg reported the use of dose-bin data obtained from the literature in order to assess the uncertainty in the estimated dose–response relation for radiation myelopathy and pneumonitis using statistical bootstrap analysis [5]. The advantage of the dose-bin method is its practical convenience, because the data needed for analysis are the number of patients and the number of occurrences of the endpoint in each dose bin. This results in less ethical concern because patient identifier data will not necessarily be revealed. Therefore, dose-bin data is another option for acquiring individual data for large-volume data sharing among institutions. However, the accuracy of using this type of data has yet to be warranted.

The present study uses radiation-induced liver disease (RILD) in liver tumour patients as a study model for describing the methodology of assessing ΔNTCP with uncertainty, using the 95% confidence interval (95% CI), between XRT and PBT. It also compares the results derived from individual patient datasets with those derived from dose-bin datasets, using various methods.

MATERIALS AND METHODS
NTCP model for RILD
RILD is a dose-limiting toxicity of liver radiotherapy and occurs with a frequency of ~5–10% when whole liver is irradiated with up to 30–35 Gy [6, 7], but a tumour requires 60–70 Gy for curative purposes. The most commonly used NTCP model for RILD is the Lyman–Kutcher–Burman (LKB) model [8, 9]. The three parameters of the LKB NTCP model are \( TD_{50} \) (the 50% tolerance dose of whole organ), \( m \) (the steepness of the dose–response curve) and \( n \) (the volume effect). To account for non-homogenous irradiation to the organ, the generalized equivalent uniform dose (gEUD) was adopted [10, 11]. The physical dose from the PBT plan, assuming an RBE of 1.1, and the fractionation schemes should be converted into gEUD as described in a supplementary document (Appendix 1).

Study scheme and virtual patient dataset generation
The study scheme is illustrated in Fig. 1. Due to the lack of a large set of DVH data and observed toxicities in liver cancer patients, we created a virtual patient dataset mimicking the Michigan data [12, 13]. In Monte Carlo fashion, a set of virtual patients in whom the statistics on mean normal liver dose (MNLD) distribution and RILD events were similar to the Michigan data was generated (REF dataset). Subsequently, the REF dataset was organized by dividing the dose level into equal intervals of 5 Gy (0–5 Gy, 5–10 Gy, 10–15 Gy, and so on) and then counting the number of RILD cases and the total number of patients in each dose level/bin (dose bin, DB, dataset) (Appendix 2).

Determination of LKB NTCP model parameters
The two types of dataset included the input data for determining the LKB NTCP model parameters using maximum likelihood estimation (MLE), where \( TD_{50} \) and \( m \) were adjusted to maximize the probabilities of predicting complications for those who experienced RILD and of predicting no complications for those who did not [14]. Subsequently, the variance \( (\sigma^2 \text{ or } \text{var}) \) and covariance \( (\text{cov}) \) of \( TD_{50} \) and \( m \) were obtained from the observed Fisher Information Matrix (FIM). The approximate 95% CI was evaluated as 1.96 standard deviations of the mean \( (\sigma \text{ or } \sqrt{\text{var}}) \), according to the central limit theorem.

Definition of the ΔNTCP function
With gEUD from the treatment plan and the estimated LKB NTCP parameters, \( TD_{50} \) and \( m \), from MLE, the ΔNTCP between XRT and PBT is given by the following function:

\[
\Delta \text{NTCP} = \text{NTCP}_{\text{XRT}} \left( g\text{EUD}_{\text{XRT}}, TD_{50}(1), m \right) - \text{NTCP}_{\text{PBT}} \left( g\text{EUD}_{\text{PBT}}, TD_{50}(1), m \right),
\]

where \( f \) is a function of ΔNTCP between XRT and PBT and \( g\text{EUD}_{\text{XRT}} \) and \( g\text{EUD}_{\text{PBT}} \) denote the normal liver dose for a certain patient for XRT and PBT, respectively. \( TD_{50} \) is the 50% tolerance dose for uniform distribution for the whole organ, and \( m \) is the steepness of the dose–response curve at \( TD_{50} \).

Determination of the ΔNTCP with uncertainty
Given the function of the ΔNTCP and the estimated \( \text{var} \) and \( \text{cov} \) matrix, the Delta method was applied [15]. Briefly, the Delta method is a standard statistical method for obtaining an approximation of the variance of a function.
where \( \partial f / \partial TD_{50} \) and \( \partial f / \partial m \) are partial differentials of \( f \) with central estimates of \( TD_{50} \) and \( m \), \( var(TD_{50}) \) and \( var(m) \) are variances of \( TD_{50} \) and \( m \), respectively, and \( cov(TD_{50}, m) \) is a covariance between \( TD_{50} \) and \( m \).

Four different algorithms for assessment of the \( \Delta NTCP \) uncertainty (95% CI) were proposed in the present study. Algorithm #1: the Delta method was applied using the REF dataset to define \( \Delta NTCP \) with a 95% CI (\( \Delta NTCP_{REF} \)) as described above. Algorithm #2: the same procedures were performed using the DB dataset (\( \Delta NTCP_{DB} \)).

Using the bootstrapping technique, newly synthesized individual datasets were generated from the DB dataset. In this case, a thousand bootstrap replicates were generated by random sampling with replacement within each dose-bin. As a result, each bootstrap replicate represented an alternative outcome in a different set of patients of the same size from the same population and contributed to different values of \( TD_{50}(1) \) and \( m \) and to their variability (\( var \) and \( cov \)). Algorithm #3: the family of parameters and variability were analysed using a method introduced by Efron [16] to identify the means of \( TD_{50}(1), var \) and \( cov \). Subsequently, the \( \Delta NTCP \) with uncertainty...
was evaluated using the Delta method (ΔNTCPBS1). Another method for assessing the uncertainty due to sampling variability was proposed by Wedenberg [5] and was used here. Algorithm #4: the 95% bootstrap CI of the ΔNTCP was estimated by analysis of the family of NTCP and ΔNTCP values (ΔNTCPBS2).

All analyses were conducted in R statistics (R Development Core Team, 2010) [17].

RESULTS

Estimated LKB NTCP parameters

From the generated REF dataset (Algorithm #1), the average MNLD among 203 patients was 29.9 Gy (range, 15.2–43.7 Gy), compared with 32 Gy (range, 14.9–44 Gy) in the Michigan study’s original data. The average MNLD was 40 Gy and 28.9 Gy for RILD and non-RILD patients, respectively, in the REF dataset and was 37 Gy and 31.3 Gy in the Michigan study. The re-estimated TD50(1) and the parameter m were 43.2 Gy (95% CI 39.1–47.3) and 0.18 (95% CI 0.11–0.24), respectively, which were convincingly similar to the original parameters, 43.3 Gy (95% CI 41.9–52.8) and 0.18 (95% CI 0.14–0.24) [12]. Figure 2 illustrates the estimated NTCP and gEUD of 203 patients with 19 RILD in the REF dataset resembling the Michigan study data (Fig. 2a), with an estimated NTCP curve with 95% CI as a function of gEUD of normal liver, considering cov(TD50, m) (Fig. 2b).

The re-estimated parameters from the DB dataset without (Algorithm #2) and with the bootstrapping technique (Algorithms #3 and 4) were compared with those from the Michigan study and the REF dataset, as shown in Table 1. Comparisons of the NTCP curves of Algorithm #1, #2 and #3 showed nearly identical results (Fig. 2c). Note that no representative NTCP curves, TD50(1) or m were identified for Algorithm #4 due to the nature of the uncertainty estimation by curve analysis.

Estimated ΔNTCP with 95% CI

From the REF dataset (Algorithm #1), the var of TD50(1) and m were 4.41 and 0.0010, respectively, which was similar to 4.32 and 0.0001.
Table 1. LKB NTCP parameters with 95% confidence intervals and their variance and covariance, according to the four proposed algorithms

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
<th>#4*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
<td>Michigan’s data</td>
<td>REF dataset</td>
<td>DB dataset</td>
<td>BS from DB dataset</td>
</tr>
<tr>
<td>LKB NTCP parameters (95% CI)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$TD_{50}(1)$, Gy</td>
<td>43.3 (42.9–52.8)</td>
<td>43.2 (39.1–47.3)</td>
<td>43.4 (39.4–47.5)</td>
<td>43.7 (38.7–48.7)</td>
</tr>
<tr>
<td>$m$</td>
<td>0.18 (0.14–0.24)</td>
<td>0.18 (0.11–0.24)</td>
<td>0.18 (0.11–0.24)</td>
<td>0.18 (0.11–0.24)</td>
</tr>
<tr>
<td>Variance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$TD_{50}(1)$</td>
<td>NA</td>
<td>4.41</td>
<td>4.32</td>
<td>6.47</td>
</tr>
<tr>
<td>$m$</td>
<td>NA</td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.0011</td>
</tr>
<tr>
<td>Covariance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA</td>
<td>0.0519</td>
<td>0.0502</td>
<td>0.0606</td>
<td>NA</td>
</tr>
</tbody>
</table>

*No representative NTCP curves, $TD_{50}(1)$, $m$ including their variability were identified for Algorithm #4 due to the nature of uncertainty estimation by curve analysis. LKB = Lyman–Kutcher–Burman, NTCP = normal tissue complication probability, CI = confidence interval, REF = reference, DB = dose-bin, BS = bootstrapping. $TD_{50}(1)$ = the 50% tolerance dose for whole organ, $m$ = steepness of the dose–response curve at $TD_{50}(1), NA = not available.

Fig. 3. The distribution of model parameters from Algorithm #1 (▪), #2 (▴) and 1000 bootstraps (+) with the mean value #3 (♦) (Fig. 3a). The correlation coefficients ($R$) between $TD_{50}$ and $m$ from Algorithms #1 (solid line), #2 (dashed line) and #3 1000 bootstrap cases (+) suggested a strong relationship between $TD_{50}$ and $m$ (Fig. 3b). $TD_{50}(1)$ = the 50% tolerance dose for whole organ, $m$ = steepness of dose–response curve at $TD_{50}(1), R = Pearson’s correlation coefficient, ALGO = algorithm.

0.0010 derived from the DB dataset (Algorithm #2). The $cov(TD_{50}, m)$ was calculated as 0.0519 and 0.502, respectively. The correlation coefficient calculated from $var$ and $cov$ showed a positive strong relationship between $TD_{50}(1)$ and $m$, with a value of 0.778 in the REF dataset and 0.768 in the DB dataset. Using the bootstrapping technique, the distribution of the model parameters from 1000 bootstrap replicates is shown in Fig. 3a and b, with a correlation coefficient of 0.8. Again, $var$ and $cov$ were not assessed in Algorithm #4.

The $\Delta$NTCP$_{REF}$, as a function of gEUD between XRT and PBT, is shown in the contour lines of the central estimate (Fig. 4a) and the 95% CI lower boundary (Fig. 4b). Each line represents the iso-$\Delta$NTCP. With the contours of $\Delta$NTCP, we can select the patient for PBT based on the confidence level. For example, in Fig. 4c, using the contour of the 95% CI lower boundary, the PBT is favoured for those with a normal liver dose of 36 Gy or more with the XRT plan, with a 95% confidence that the $\Delta$NTCP is 10% or higher. In contrast, using the contours of the central estimate, PBT likely provides a potential benefit for patients who receive a normal liver dose of 33 Gy or more with the XRT plan, with less confidence. Thus, the use of a 95% CI lower boundary contour is more conservative for patient selection for PBT than is using the central estimate contour. However, in the area of 33–36 Gy, the
selection of the treatment modality should be determined by considering the trade-offs between clinical benefits and socio-economic aspects.

The contours of the 95% CI lower boundary of the ΔNTCP at 10% among four proposed algorithms are compared in Fig. 5. Compared with the ΔNTCPREF (Algorithm #1), the contour of the ΔNTCPDB (Algorithm #2) was nearly indistinguishable, as were the ΔNTCPBS1 (Algorithm #3) and the ΔNTCPBS2 (Algorithm #4). This was true, in spite of the varying percentage differences of the ΔNTCP. However, the contours from the bootstrapping techniques yielded little differences. This might be due to the larger cov found in the bootstrap dataset.

**DISCUSSION**

Considering the Michigan study’s parameters and the number of patients reported in the literature [12], the REF dataset was a good representation of the Michigan data, despite a small difference in the MNLD. The re-identified model parameters were remarkably similar to the original ones. However, variant 95% CI values were observed due to the different methods used to obtain the uncertainty, i.e., the profile likelihood method used in the Michigan study versus the variance-based method based on the central limit theorem used in our study. With virtual simulated gEUD data from 203 patients, the ΔNTCP with 95% CI between the treatment modalities could be determined using the Delta method. The 95% CI lower boundary of the ΔNTCP provides a more conservative threshold for selecting patients for PBT compared with the central estimate. Our results also demonstrated that the LKB NTCP model parameters and variability derived from the dose-bin dataset were very similar to those from the individual patient dataset. Additionally, the contour of the ΔNTCPDB was very similar to the one of the ΔNTCPREF.

Liver cancer is common in Eastern and Southeastern Asia [18]. Data from Japan on comprehensive cancer statistics show that it is the fifth most common cancer, with an estimated incidence of 45,100 cases, and liver cancer was the fifth most common cause of cancer deaths in 2015 [19]. PBT is an effective treatment modality for both primary and secondary liver malignancies [20–25]. A prospective Phase II clinical study demonstrated the efficacy and feasibility of PBT in primary liver tumours [26]. Although a randomized controlled trial (RCT) comparing PBT with standard XRT is expected in the future, there are several obstacles impeding the...
was a more con-
and co
= was slightly different in our Michigan-resembling
by Hokkaido University user
Downloaded from https://academic.oup.com/jrr/article-abstract/59/suppl_1/i50/4930790
affected by the uncertainty in the NTCP models [78x624] the accuracy of the MBA on patient selection for PBT is largely
able NTCP models in head and neck cancers and concluded that
distribution (mean and CI) of the model coef-
patients with 13 tumours [78x371] i56
 considered.
NTCP curves would be better estimated when
the parameter
NTCP model is currently more appealing due to its feasibility in
development and implementation for allocating the best treatment
modality to individual patients [1, 2, 29].
Bijman et al. assessed the model uncertainty using a probability
distribution (mean and CI) of the model coefficients from multivari-
able NTCP models in head and neck cancers and concluded that
accuracy of the MBA for patient selection for PBT is largely
affected by the uncertainty in the NTCP models [4]. In contrast,
the uncertainty of the LKB NTCP model for RILD relies on the
model parameters’ variability, var and cov. Due to the high correl-
ation between TD50 and m, we assumed that the uncertainty of the
NTCP curves would be better estimated when cov(TD50, m) was
considered.
The mean value of the NTCP was generally applied to obtain the
\( \Delta NTCP \) [1–3]. In the present study, we showed that the lower
boundary of the 95% CI, considering var and cov, was a more con-
servative threshold of \( \Delta NTCP \) for decision-making regarding the
use of PBT. In clinical practice, for a certain patient, a treatment
plan comparison between XRT and PBT is performed, based on the
dosimetric difference of normal liver DVHs. When using our iso-
\( \Delta NTCP \) curve (Fig. 4), if the gEUD falls within the area where the
\( \Delta NTCP \) is more than the predefined \( \Delta NTCP \) threshold (the area
to the right of and beneath the contour), a PBT can be chosen for a
particular patient. This 95% CI lower boundary method conserva-
tively selects those patients who potentially could benefit from PBT.
Thus, the overuse of PBT can be prevented. However, the prede-
\( \Delta NTCP \) should be appropriately determined based on clinical
outcomes and cost-effectiveness studies.
Toramatsu et al. performed a dosimetric comparison between
spot-scanning proton therapy (SSPT) and IMRT in 10 HCC
patients with 13 tumours [30]. We obtained the gEUDs, or mean frac-
tion size equivalent doses (FEDs), from the XRT and PBT plan from
table 2 in Toramatsu et al.’s publication and applied them to our iso-
\( \Delta NTCP \) contours. According to Toramatsu’s study, tumours with a
nominal diameter of \( >6.3 \) cm \((8/13 \) tumours) had an average risk of
RILD of 6.2% for SSPT and 94.5% for IMRT, corresponding with our
10% iso-\( \Delta NTCP \) 95% CI lower boundary contour (Fig. 6). However,
the parameter \( n \) was slightly different in our Michigan-resembling
NTCP curve \((n = 1.1)\) and Toramatsu’s publication \((n = 0.97)\).
Based on the types of available data, our study suggests that the
DB data can be utilized and achieves the same results as individual
patient data, with more convenience in data collection and sharing.
Algorithm #2 constantly provided results similar to those achieved
with Algorithm #1, whereas Algorithms #3 and #4 were affected by
statistical instability and consumed a large amount of time.
Due to the lack of individual patient data, we created a virtual
simulated patient dataset based on the Michigan study data, in
which the fractionation scheme was unique \((1.5 \) Gy twice daily) and
concurrent chemotherapy was administered. In this study, the LKB
model was used to create the map, but there were several types of
NTCP models assuming different equations and parameters and
resulting in different NTCP values and \( \Delta NTCP \) contours. Thus,
our iso-\( \Delta NTCP \) contours should be interpreted cautiously. Further-
more, the NTCP model derived from the Michigan study was dedi-
cated to patients with normal liver function treated with XRT, and

Fig. 6. Fraction size equivalent doses (FEDs) of 13 liver
tumours from Toramatsu’s data (filled squares) are plotted
with the contour of the iso-\( \Delta NTCP \) 95% CI lower boundary
derived from Algorithms #1 (dashed line) and #2 (solid
line). The small numbers above denote tumour size in
centimeters. gEUD = generalized equivalent uniform dose,
XRT = X-ray therapy, PBT = proton beam therapy, ALGO =
algorithm.
the PBT is theoretically useful for those with impaired hepatic func-
tion \((\text{Child–Pugh B and C})\), as suggested by Dawson [31]. As a
result, the NTCP model needs to be prospectively developed and
needs to include these patient subgroups in order to develop an
accurate and reliable iso-\( \Delta NTCP \) decision-making map. In the
future, clinical validation studies and cost-effectiveness analyses are
expected to select patients who potentially could benefit from PBT.
In conclusion, the methodology presented in this paper relies on
systematic statistical considerations of the 95% CI based on the
Delta method and, considering the variance–covariance matrix, can
be applied to other types of NTCP models and tumours. This
might ultimately establish a guideline for properly selecting patients
for proton therapy.

SUPPLEMENTARY DATA
Supplementary data is available at Journal of Radiation Research
online.

CONFLICT OF INTEREST
The authors state that there are no conflicts of interest.

REFERENCES
1. Langendijk JA, Lambin P, De Ruysscher D et al. Selection of patients
for radiotherapy with protons aiming at reduction of side effects: the


