Journal of Physics: Conference Series **489** (2014) 012087 doi:10.1088/1742-6596/489/1/012087

An MLE method for finding LKB NTCP model parameters using Monte Carlo uncertainty estimates

Martin CAROLAN[1,2], Brad OBORN[1,2], Kerwyn FOO [3], Annette HAWORTH[4], Sarah GULLIFORD[5] and Martin EBERT[6,7]

[1]Illawarra Cancer Care Centre, Wollongong Hospital, Wollongong NSW 2500, Australia

[2]Centre for Medical Radiation Physics, University of Wollongong, Wollongong NSW 2522, Australia

[3]Department of Radiation Oncology, Royal Prince Alfred Hospital, Camperdown, NSW 2050, Australia

[4]Peter McCallum Cancer Centre, Melbourne, Vic, Australia

[5]Joint Dept of Physics, Institute of Cancer Research and Royal Marsden National Health Service Foundation Trust, Sutton, United Kingdom

[6]Sir Charles Gairdner Hospital, Nedlands, WA, 6009, Australia
[7]School of Physics, University of Western Australia, Crawley, WA 6009, Australia.

E-mail: martin.carolan@sesiahs.health.nsw.gov.au

Abstract.
The aims of this work were to establish a program to fit NTCP models to clinical data with multiple toxicity endpoints, to test the method using a realistic test dataset, to compare three methods for estimating confidence intervals for the fitted parameters and to characterise the speed and performance of the program.

1. Introduction

The QUANTEC reports [1] have summarised and reported data on radiation induced complication probabilities in multiple organs. The QUANTEC series emphasised the importance of collecting dosimetry and outcome data to enable more accurate modelling of normal tissue complication probability (NTCP). This data has mainly been collected in the context of clinical trials in the past but ideally clinics would record detailed and accurate data for all patients during routine treatment operations allowing a "data mining" approach to development of more accurate NTCP models [2]. Various NTCP models exist but one of the most commonly applied models is the Lyman, Kutcher, Burman (LKB) model [3]. The LKB model describes the NTCP in terms of three model parameters and the dose received by the organ of interest. When the model parameters are known and the dose volume histogram data is available the calculation of the NTCP is trivial and can be performed in a few seconds using a spreadsheet. Determining the appropriate values of the model parameters from a large set of patient dose volume histograms and observed toxicities is significantly more computationally intensive. The maximum likelihood estimation (MLE) method is used to identify the set of parameter values which best describes the observed clinical outcomes. The estimated LKB model parameters are drawn from some distribution of possible values that describe the patient group and it is necessary to determine the confidence interval around these values. There are several methods available to determine

the confidence intervals including the profile likelihood method as well as Monte Carlo based techniques such as parametric and non parametric bootstrapping. This paper describes the design and testing of a program to determine LKB model parameters from a number of test data sets and compares parameter confidence intervals derived using parametric and non parametric bootstrapping.

2. Methods

2.1. Lyman - Kutcher - Burman Model

The LKB model [3] is used to generate a normal tissue complication probability based on the dose received by the normal tissue of interest. The dose volume histogram (DVH) for the normal tissue is collapsed into a single dose parameter referred to as the Equivalent Uniform Dose (EUD). Conceptually the EUD may be considered as the uniform dose to a particular organ that would yield the same outcome or toxicity endpoint. The Equivalent Uniform Dose is calculated as follows:

$$EUD = \left[\frac{1}{N_{voxels}} \sum_{i=1}^{N_{voxels}} d_i^a \right]^{\frac{1}{a}} \tag{1}$$

The dose in each voxel of the organ is raised to the power of a (an alternative representation uses $n = 1/a$) and summed over all voxels. The parameter a represents the seriality of the tissue response to radiation. When a is large the overall EUD is dominated by the hottest voxels (serial type response). When a is small the EUD is more influenced by the total volume irradiated (parallel type response). The NTCP is then given by:

$$NTCP = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{\frac{-x^2}{2}} dx \qquad where: t = \frac{EUD - D_{50}}{mD_{50}} \tag{2}$$

Where D_{50} represents the dose to the organ that results in a 50% chance of a particular complication and m defines the slope of the assumed sigmoid dose response curve of the tissue.

2.2. Finding a, m and D_{50} using Maximum Likelihood Estimation

Given a sample of patients where the DVHs for the organ at risk of complication are known and the observed incidence of a particular complication is recorded it is possible to make an estimate of a, m and D_{50}. To achieve this NTCP values for each patient in the population are calculated for a range of a, m and D_{50} values. For each combination of a, m and D_{50} the NTCPs for each patient are compared to the observed toxicities and the log-likelihood (LLH) is calculated [4]. The log likelihood is a measure of agreement between the observed toxicities and the calculated NTCPs and is found as follows:

$$LLH = \sum_{tox=1} \ln(NTCP(a, m, D_{50})) + \sum_{tox=0} \ln(1 - NTCP(a, m, D_{50})) \tag{3}$$

The first term is summed over all patients where a complication is observed and the second is summed over all patients where no complication is observed. The combination of a, m and D_{50} which yields the maximum value of the log-likelihood is the parameter set which best fits the patient data being used. The search for the maximum likelihood can be approached with a smart optimisation algorithm or alternatively by brute force searching through a large grid parameter space. In this work we took the brute force approach.

2.3. Multiple Complication Endpoints
Once suitably large arrays of a, m, D_{50} and NTCP are calculated these are independent of the actual observed endpoint and can be used to find the MLE of the parameters for any observed complications. This calculation efficiency increases with the number of end points fitted.

2.4. Finding the Confidence Intervals
Confidence intervals are commonly determined using the profile likelihood method. This involves finding the maximum log-likelihood and then varying each parameter until the log-likelihood is decreased by an amount equal to half the critical value of the $\chi^2(1)$ distribution at the desired significance level. The search is repeated for each parameter and this yields the respective confidence intervals [5]. A more general but computationally intensive approach also commonly used in NTCP modelling is non-parametric bootstrapping [6]. This involves randomly selecting a population of patients from the data set (with replacement) and then finding the a, m, and D_{50} again. This is repeated many times (e.g. 1000 - 2000) and yields a distribution of a, m, and D_{50} values. The confidence interval is determined by looking at the value of each parameter at the percentile cut-off for the confidence interval required in each distribution. This approach relies on the intrinsic distribution of the patient sample itself. It is robust and the main potential weakness is if the patient group does not adequately cover the whole parameters space. An alternative approach is to use parametric bootstrapping. This assumes that the fitted model is a good representation of the data (and the real population it represents). Therefore a synthesised patient population can be derived from the model rather than the original patient data set. In practice this is achieved by calculating the NTCP for each patient in the original data set using the optimised a, m, and D_{50} values. To generate a new patient population, a random number (r_n) between 0 and 1 is generated for each patient. If $r_n < NTCP$ patients are assigned a complication. If $r_n > NTCP$ the patients are assigned no complication. The values of a, m, and D_{50} for the new synthesised patient population are found using MLE. This process is repeated many times (1000-2000) and the confidence intervals are determined from the resulting distributions. The parametric approach relies on the model being a good fit to the true patient population.

2.5. Implementation
The MLE optimisation and three different confidence interval estimation approaches were programmed using MATLAB (version 7.12.0, The Mathworks Inc., 2011). The code was run on a HP server with 64 cores (four 16 core CPUs) running at 2.6 GHz and with 256 GB of RAM. The array containing the grid of a, m, D_{50} and NTCP for each patient was calculated once. All subsequent MLE searches were performed against this array. Since this array is accessed 1000's of times in the bootstrapping routines using a large precalculated grid is advantageous. The search grid was defined so that a had a range from 10^{-3} to 10^{2} in 61 bins equally spaced on a logarithmic scale. The dose parameter D_{50} spanned from 0-150 Gy in 1 Gy increments. The NTCP slope parameter m spanned from 0.01 to 2 in 47 bins. To perform the non parametric bootstrapping 2000 re-sampled populations were broken up into 50 batches of 40 populations thereby using 50 of the 64 cores on the server. The parametric bootstrapping was performed in a similar way with 2000 synthesised patient populations solved across 50 cores.

2.6. Evaluation with a Test Data Set
Three test data sets were based on DVHs and toxicities from 738 patients in the TROG 03.04 RADAR clinical trial [7, 8]. For the purposes of the current work the data set was treated simply as a realistic distribution of DVHs. The DVHs for all patients were used in a single data set without consideration of any other details related to the trial. The test set model parameters derived and used in this work do not have any clinical significance beyond the testing described

here and readers are referred to references [7, 8] for more details of the RADAR study itself. Three separate complication endpoints were chosen giving a range of parameter values and numbers of patients with the complication.

3. Results

The test data sets consisted of anal canal and anorectal dose volume histograms from 738 patients in the RADAR data set. All patients for which complete data was available were included. The data set therefore incorporates a reasonably heterogeneous set of patients from all arms of the study with prescribed doses ranging from 66 to 74 Gy. Using various techniques. The complication incidence for the three test data sets was 126/738, 309/738 and 126/138 for test sets 1, 2 and 3 respectively.

3.1. Confidence Intervals

In all cases the confidence interval derived using the likelihood profile method is smaller than either of the bootstrapping methods for deriving confidence intervals. The confidence interval derived from the non parametric bootstrap was always larger than the confidence interval based on the parametric bootstrap. This trend is expected since the likelihood profile confidence interval represents the MLE fit to the specific data set. With the parametric bootstrap the possibility of different outcomes for each member of the patient test population is exercised but with the inherent assumption that the NTCP model accurately describes the distribution of the actual data. Whereas the non parametric model which re-samples the existing patient population does not assume the model accurately describes the patient distributions. Examples of the a, m and D_{50} distributions found for the two different bootstrap methods are shown in figures below.

Table 1. Model parameter values and confidence intervals (68%) for each calculated by three different methods each of the test data sets.

	a	D_{50}	m
Test Set 1	0.6	69	0.6
likelyhood profile CI	0.001-0.7	66-79	0.6-0.65
parametric CI	0.034-0.2	62-81.5	0.525-0.7
non parameteric CI	0.034-3.5	61-86.5	0.47-0.71
Test Set 2	3.2	55	0.55
likelyhood profile CI	2-5	52-59	0.45-0.85
parametric CI	1.46-6.24	50.3-60.6	0.41-0.94
non parameteric CI	0.88-7.24	47.6-61.4	0.37-1.03
Test Set 3	3.2	77	0.4
likelyhood profile CI	2-5	72-88	0.3-0.5
parametric CI	1.17-7.15	70.7-93.5	0.32-0.61
non parameteric CI	1.21-11.73	69.9-94.7	0.26-0.61

3.2. Computation Time

Our test data sets had 738 patients and the MLE search was performed over a very modest grid of a, m, D_{50} values of dimensions $(61 \times 150 \times 47)$. To construct the array of EUD and NTCPs required for each patient with this search grid takes approximately 30 minutes on our hardware. A single MLE solution of a, m, D_{50} for all three of the test data sets took 150 seconds on a single core. Therefore to perform a MLE solution for 2000 patient populations takes approximately 3×10^5 seconds or 83 hours. When this is run on 50 cores simultaneously (without memory sharing) the calculations can be completed in approximately 1 h 40 minutes.

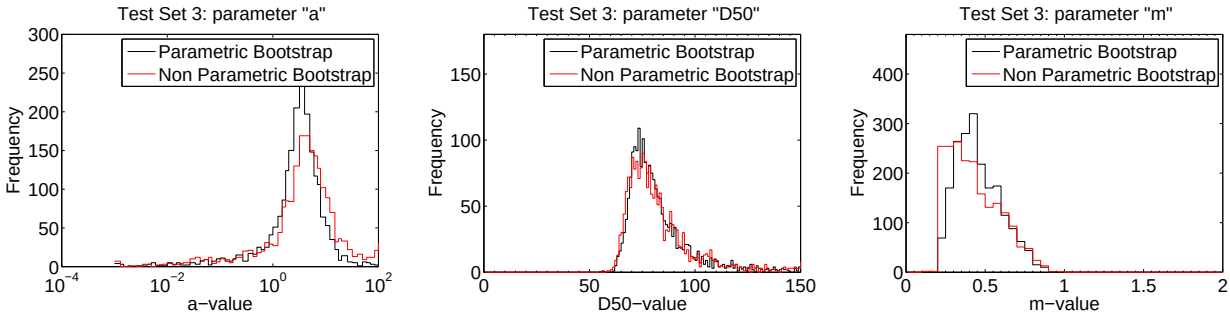

Figure 1. Parametric and Non Parametric bootstrapping for the parameters "a", "D_{50}" and "m".

There was no difference observed in the computing time (150 ± 5 seconds per iteration) required for the parametric versus the non parametric bootstrapping techniques so in our implementation neither has any advantage over the other in this respect. At these speeds of execution and relying on multiple cores for processing it is possible to analyse a data set with approximately 750 patients and derive NTCP parameters for 20 different complication end points in an overnight batch taking approximately 12 hours.

4. Discussion

4.1. Grid search considerations

Determination of LKB parameters calculating the EUD and NTCP on the fly for a large number of patients and end points is very inefficient. A pre-calculated grid is preferred. The span of the grid is important so that the search space contains any real MLE maxima. Failure to meet this requirement is indicated by overpopulated bins at the extremes of the bootstrap distributions. One shortcoming of our implementation of the "brute force" grid search technique is that it does not easily lend itself to adaptive grid resizing as this would require further NTCP calculations on the fly as the MLE converges on a maximum value. A smarter (no grid) optimisation procedure could give better performance so long as it converges quickly (i.e. the total number of NTCP calculations for all bootstrapping and all endpoints is significantly less than the 4.3×10^5 array entries used for the grid search).

4.2. Parametric or non parametric bootstrap?

This work shows that the choice of either parametric bootstrapping or non-parametric bootstrapping for confidence intervals in NTCP models is probably not critical in most cases. The non parametric approach is used in most publications related to confidence intervals for NTCP models. This assumes the sample of patients in the study group is representative of the wider population to whom the model would be applied. The bootstrapping procedure should give a good indication of the distribution the parameter could have for that population. In contrast to this the parametric approach results in a distribution of parameter values reflecting the model prediction for these values not the actual sample set distribution. This will give narrower confidence intervals than the more pessimistic confidence intervals from the non parametric bootstrap. Derivation of the confidence intervals was based on a simple quantile approach however other approaches to determining the confidence intervals are described in the statistics literature [6]. These are recommended as more rigorous than our simple approach and should be considered for future implementation. A parametric approach is invaluable for predicting distributions of NTCPs for a set of patients and comparing how well different models fit the patient data [9, 10]. Either technique will provide a more conservative estimate of the confidence

intervals for an NTCP model than the profile likelihood calculation and in our implementation there was virtually no difference in the computing time.

5. Conclusions

This simple exercise has demonstrated that in a small number of test data sets parametric and non parametric bootstrapping can both be used to generate comparable confidence intervals for NTCP model parameters. Both approaches give more information about parameter distributions than the simple but less intensive likelihood profile method. There is no difference in calculation time in our implementation of the calculations. Further refinement including implementation of an adaptive grid spacing would improve accuracy of the MLE optimisation but this is difficult to incorporate into a simple grid search algorithm without speed penalties. Additional functionalities to allow cross validation of fitted models and development of well characterised data sets as suggested by other authors [11] should also be implemented.

6. Acknowledgements

This work is supported by NHMRC Project Grant 1006447. We thank Dr Anthony Carolan for useful discussions on parametric and non parametric bootstrapping.

References

[1] Marks, L. B., et al. (2010). IJROBP, 76(3 Suppl), S10–9. doi:10.1016/j.ijrobp.2009.07.1754
[2] Deasy, J. O., et al. (2010). . IJROBP, 76(3 Suppl), S151–4. doi:10.1016/j.ijrobp.2009.06.094
[3] Stewart R.D., Li, X.I. (2007), Med Phys 34(10):3739-3751.
[4] Gulliford, S. L., et al, (2012). Radiother and Oncol, 102(3), 347-51. doi:10.1016/j.radonc.2011.10.022
[5] Stryhn, H., & Christensen, J. . In Proceedings of 10th International Symposium Veterinary Epidemiology and Economics, 2003. p208.
[6] Carpenter, J., & Bithell, J. (2000). . Statistics in medicine, (August 1999), 1141-1164. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-0258(20000515)19:9<1141::AID-SIM479>3.0.CO;2-F/abstract
[7] Denham, J. W., et al. (2012). Radiother and Oncol, 105(2), 184–92. doi:10.1016/j.radonc.2012.09.018
[8] Kearvell, R, et al. J Med Imaging Radiat Oncol. 2013 Apr;57(2):247-57. doi: 10.1111/1754-9485.12025. Epub 2013 Jan 7
[9] Seppenwoolde, Y. V. S., et al. (2003) 55(3), 724–735. doi:10.1016/S0360-3016(02)03986-X
[10] Schilstra, C., & Meertens, H. (2001). .IJROBP, 50(1), 147–58. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11316558
[11] Xu, C.-J., Van der Schaaf, A., et al (2012). IJROBP, 84(1), e123–9. doi:10.1016/j.ijrobp.2012.02.022