1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Factors influencing the robustness of P-value measurements in CT texture prognosis studies

Sarah McQuaid^{1,2}, James Scuffham^{1,3}, Sheaka Alobaidli³, Vineet Prakash^{1,4,5}, Veni Ezhil¹, Andrew Nisbet¹, Christopher South¹, Philip Evans³

¹Royal Surrey County Hospital NHS Foundation Trust, Guildford, UK

²University College London Hospitals NHS Foundation Trust, London, UK

³University of Surrey, Guildford, UK

⁴Ashford and St. Peter's Hospitals NHS Trust, Chertsey, UK

⁵Guildford Diagnostic Imaging, Guildford, UK

Abstract

Aim: Several studies have recently reported on the value of CT texture analysis in predicting survival, although the topic remains controversial, with further validation needed in order to consolidate the evidence base. The aim of this study was to investigate the effect of varying the input parameters in Kaplan-Meier analysis, to determine whether the resulting P-value can be considered to be a robust indicator of the parameter's prognostic potential. Methods: A retrospective analysis of the CT-based normalised entropy of 51 patients with lung cancer was performed and overall survival data for these patients were collected. A normalised entropy cut-off was chosen to split the patient cohort into two groups and log-rank testing was performed to assess the survival difference of the two groups. This was repeated for varying normalised entropy cut-offs and varying follow-up periods. Our findings were also compared with previously published results to assess robustness of this parameter in a multi-centre patient cohort. Results: The P-value was found to be highly sensitive to the choice of cut-off value, with small changes in cut-off producing substantial changes in P. The P-value was also sensitive to follow-up period, with particularly noisy results at short follow-up periods. Using matched conditions to previously published results, a P-value of 0.162 was obtained. Conclusions: Survival analysis results can be highly sensitive to the choice in texture cut-off value in dichotomising patients, which should be taken into account when performing such studies to avoid reporting false positive results. Short follow-up periods also produce unstable results and should therefore be avoided to ensure the results produced are reproducible. Previous published findings that indicated the prognostic value of normalised entropy were not replicated here, but further studies with larger patient numbers would be required to determine the cause of the different outcomes.

Introduction

The field of radiomics has received substantial attention in recent years, with many publications outlining the potential benefits of applying texture analysis to diagnostic images to improve predictions of survival or time to disease progression (Alobaidli et al 2014, Cheng et al 2013, Cook et al 2013, Ganeshan et al 2012a, Ganeshan et al 2012b, Goh et al 2011, Mayr 2012, Miles et al 2009, Ng et al 2013, Win et al 2013, Yip et al 2014, Zhang et al 2013), improving predictions of treatment response (Cook et al 2013, Mattonen et al 2014), improving disease classification (Eliat et al 2012, Ganeshan et al 2009, Vasconcelos et al 2010, Xu et al 2006) and providing useful information in radiotherapy treatment planning (Alobaidli et al 2014, Yu et al 2009). Despite this, there are many aspects of the field that still need to be clarified, due to conflicting results in some publications, different methodology used in different studies and a lack of fundamental understanding of the

relationship between image texture and the underlying biology from which it arises (Buvat et al 2015, Chalkidou et al 2015). Combined with the difficulty in determining which texture parameters have a genuine relationship with the outcome being tested and which are false positives, this has led some authors to question the validity of the evidence base and whether or not there is currently sufficient evidence of its efficacy in the clinical setting (Chalkidou et al 2015).

One of the contributory factors in the over-reporting of false positives is the use of 'optimised cut-off values' when dichotomising patients for Kaplan-Meier analysis and significance testing (Altman et al 1994). In this approach, the cut-off value of the texture metric used to split the patient cohort into two groups is chosen such that the P-value in a log rank test to compare the outcomes of the two groups is minimised. In other words the difference in the outcome measure (e.g. survival) between the two groups is maximised by choosing how the groups are formed. Since this approach requires multiple significance tests to be performed with only the best result being reported, it is associated with a high false-positive rate (Altman et al 1994, Hilsenbeck et al 1992). While this is an appropriate method for an initial exploratory study to identify which texture parameters may be of interest, further validation studies are required before it can be asserted that the texture parameters are genuinely related to outcome. To date, few validation studies are available in the literature (Chalkidou et al, 2015 looked at 15 recent publications on texture analysis and only 3 of these had incorporated validation steps into the study design) and hence it is unclear which texture parameters are of clinical interest and which appear to have a relationship with outcome purely by chance. Although strategies for minimising false positive results are available (Benjamini and Hochberg 1995, Holm 1979) these have not been in widespread use in texture analysis studies to date (Chalkidou et al 2015). Together with the lack of knowledge on how stable the results obtained are (with respect to choice in cut-off value or follow-up time), this has led to uncertainty regarding the true prognostic power of texture analysis metrics. It is also unclear how sensitive the results obtained are to the choice in cut-off value, which is an important consideration in the clinical implementation and the efficacy of the technique in the wider clinical setting.

Another factor that can affect results is the duration of follow-up and how many events are observed in that period. A short follow-up period is likely to result in few events (deaths in the case of survival studies), which could limit the robustness of the result obtained. It is recommended that at least 10 to 15 events are observed per predictor (Babak 2004, Peduzzi et al 1996) in order to have confidence in the resulting P-value, but this is commonly not achieved, due to the difficulties associated with obtaining a large sample and with sufficient follow-up. If the P-value were sensitive to the follow-up period, whereby recording an additional event produced a substantial change in the result, this would suggest that false-positives were likely and that further follow-up data were required to obtain a robust result. However, normally results are only obtained for a single follow-up period and so the sensitivity of the P-value to follow-up is not known. The purpose of this study was therefore to investigate the effect of both the choice of cut-off value and the follow-up period to determine how these influence the result and to consequently determine how robust the P-value is to these choices. Furthermore, the purpose was also to inform future texture analysis studies, in helping to interpret results and to determine the likelihood of a robust, reproducible result having been obtained. To date, little cross-centre validation of texture analysis metrics for prognosis has been undertaken and so their reproducibility in different patient groups has not been established. If texture parameters only yield prognostic information under certain circumstances (e.g. for a particular disease stage or histology), this could limit their use to smaller, specific cohorts of

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

patients, rather than being generally applicable to all lung cancer patients, for example. Comparing our results with a previously validated metric could therefore provide some insight into how robust and reproducible the metric is across a wider set of patient characteristics, across different centres.

The influence of parameter choice on the result obtained was investigated by collecting patient data and assessing the association between patient survival and a CT-based normalised entropy calculation with different cut-off values and follow-up times. Normalised entropy (where a ratio is calculated between the entropy measures obtained with different sized filters) was chosen as the metric of interest since Win et al (2013) have reported that this parameter has prognostic value for NSCLC patients. By evaluating a parameter that has previously been shown to have potential as a prognostic tool, a validation of this parameter for a similar cohort of patients was possible. A comparison against their results was enabled by matching the normalised entropy cut-off they previously found to be optimal and their follow-up period.

Other parameters were not considered here, since the aim was not to perform an exploratory study looking into which texture parameters have prognostic potential; instead the focus was on the stability of results obtained and on whether previously obtained positive findings can be confirmed. The final part of the study involved the generation of simulated survival data to allow further investigation into the effects of changing the follow-up period and the conditions required for a robust result.

Methods

Patient data

Patient cohort

For this study, 72 patients who underwent IMRT for lung cancer between 2009 and 2012 were retrospectively identified, however only 51 of these patients could be fully processed due to the tumour being too small for a calculation of normalised entropy in the remaining 21 patients (further information on the processing methods is given below). Details of the patient cohort are given in Table 1. The clinical records of these patients were reviewed to determine their survival status. Each patient received a diagnostic PET-CT scan on a GE Discovery system (GE Healthcare, Buckinghamshire, UK) prior to starting radiotherapy; the CT components of these scans were retrieved for texture analysis.

Table 1. Clinical details of the patient cohort used in the texture analysis study

Number of patients:	51 (M: 36, F 15)	
Age at time of scan:	Mean: 67.9 y, Range: 45.9 – 87.5 y	
Histology:	Squamous cell	28
	Adenocarcinoma	12
	NSCLC of non-specific type	7
	Small-cell	2
	Histology information not available	2
Staging:	IB	3
	IIA	2

	IIB	7
	IIIA	22
	IIIB	11
	IV	4
	Staging information not available	2
Follow-up:	Survival range	3.0 – 62.2 months
	Median survival	19.6 months
	Patients alive at the end of the study	11 / 51 (21.6 %)

Imaging data and texture analysis

All CT scans were acquired with a peak tube voltage of 140 kVp and a current of 100 mA, and were reconstructed onto an image matrix in which each slice contained 512x512 pixels with a pixel size of 0.98x0.98mm. The slice thickness was 5mm. Regions of interest were manually drawn around the tumours of each patient by a single experienced radiologist (VP). The CT slice showing the largest area of tumour was then selected for texture analysis, which was conducted using a proprietary software package (Texrad, Cambridge, UK). This software processes the CT data within the region-of-interest; the region is eroded and all pixel values below -50 HU are excluded to ensure that no normal lung tissue is included within the processed data. The sizes of these processed regions ranged from 3.8 cm² to 45.7 cm², with a mean of 12.6 cm². (Note that in the original set of 72 patients, the sizes ranged from 1.4 to 45.7 cm², with a mean of 9.6 cm², but the smallest tumours could not be processed using TexRad and hence were excluded.)

hA Laplacian of a Gaussian band-pass filter was then applied to the processed region with two different widths, corresponding to medium and coarse image texture. These filters have widths of 6 and 12 pixels respectively. Some tumours were too small to apply the filters to them and hence could not be processed in TexRad; of the original 72 datasets, 70 were processed using the medium filter, but due to the larger filter size, only 51 could be processed using the coarse filter. The overall texture in the tumour region of interest was then quantified for each filter width using entropy derived from the histogram of intensity values. The normalised entropy (medium/coarse scale) was then calculated for the 51 patient datasets, which is the same approach used by Win et al (2013). (Further details of the Win et al study are given in the 'Comparison against Win et al results' section, below.)

Statistical analysis

Survival analysis was conducted using Matlab (MATLAB R2015a, The MathWorks Inc., Natick, MA). A cut-off value of normalised entropy was selected, using which the patient cohort was split into low and high normalised entropy groups. A follow-up time was then selected, and survival curves of the two groups were generated using a Kaplan-Meier approach as far as the chosen follow-up time. The statistical significance of the differences between the Kaplan-Meier survival curves was assessed using the log-rank test, with the null hypothesis of no difference between the two groups. This was repeated for 50 different cut-off values of normalised entropy so that different splits in the two patient groups were obtained each time, allowing investigation of the P-value as a function of normalised entropy cut-off. In this way the full range of normalised entropy values for these patients were used as cut-off values for the survival analysis.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

The effect of changing follow-up period was also investigated using a similar approach. For a given normalised entropy cut-off, a P-value was calculated for every time point at which an event occurred (a patient death or follow-up ceased). This resulted in 51 P-values corresponding with the 51 survival / follow-up times. A 2D array of P-values was built up by evaluating P for all follow-up times and normalised entropy cut-offs described here. Note that although multiple P-value calculations were performed, this was to assess what P-value would have been obtained if a study had been performed with those particular conditions, rather than a direct assessment of the statistical significance of each assessment. Therefore P-value corrections for multiple testing were not performed as this was not the subject of this study.

Comparison against Win et al results

Win et al assessed the prognostic value of CT normalised entropy of NSCLC patients, also using TexRad as described in Win et al (2013). Their study also used the CT component of a PET-CT acquisition acquired using a GE Discovery, with acquisition parameters of 140 kVp and 80 mA (compared with 140 kVp and 100 mA in this study). Their pixel size was also 0.98 x 0.98 mm with a slice thickness of 3.75 mm (Ganeshan et al 2012b) (compared with 0.98 x 0.98 x 5 mm voxels in this study). As used here, Win et al calculated the normalised entropy by taking the ratio of the entropy using the medium filter width (6 pixels) to the entropy calculated using the coarse filter (12 pixels) and assessed whether this metric was associated with overall survival. Using a feasibility dataset of 54 patients, they employed an optimised cut-off approach (determining the cut-off that produces the best separation in survival curves between the two groups), which was found with a normalised entropy of 1.233. This cut-off was subsequently applied to a validation dataset of 66 patients, from which a P-value of < 0.001 was obtained. The high normalised entropy group (> 1.233) were found to have a longer mean survival (39.7 months) compared with the low normalised entropy group (< 1.233, 18.2 months).

Our P-value calculation to replicate the Win et al validation study analysis was therefore performed with a normalised entropy cut-off of 1.233 and an equivalent follow-up time of 40 months.

Simulated Data

Simulated survival curves using a pseudo-random number generator in Matlab were generated in order to illustrate the problems of false positive and false negative results in the context of survival analysis. Unlike real patient data, it is possible to do this with complete follow-up (so that no censoring is required) and with a known underlying survival distribution, so that results produced can be assessed in terms of whether it is a genuine finding, a false-positive or a false-negative. The aim was therefore to produce Kaplan-Meier curves that had a similar appearance to that produced from the patient data, allowing the potential scenarios leading to that particular curve shape to be assessed.

Survival data were generated for 51 cases (to match the patient cohort size), split into 2 groups of 26 and 25 cases respectively. Kaplan-Meier analysis was initially performed for the scenario where the two groups had identical survival functions; both were defined by a normal distribution with a mean of 50 months and a standard deviation of 20 months, to produce similar Kaplan-Meier curves to

those produced by the real patient data. As with the patient data, P-values were then calculated as a function of follow-up time using the log-rank method, to produce 50 P-values in total.

A second simulation was performed where the mean survival of group 1 was longer than for group 2; again, both groups' survival patterns were taken from normal distributions, but in this case group 2's mean survival was shortened to 45 months. (The standard deviation remained fixed at 20 months). In both simulations, the aim was to generate a case where the P-value is < 0.05 at an intermediate follow-up time, but is > 0.05 with the longest follow-up (as is the case for the patient data), in order to illustrate the issues with these specific cases. Due to the variability of the results obtained using a random number generator, different shaped Kaplan-Meier curves are produced every time the simulation is run. Therefore, repeated runs were made, which were stopped when the desired characteristics were achieved. The first simulation was run twice to achieve this result, and the second was run 3 times; only the output from the final run is presented here.

Results

Patient data: varying input parameters

The result of changing the normalised entropy cut-off and the follow-up period on the subsequent P-value is shown in Figure 1. The plots in Figures 2 and 3 are single curves extracted from this surface, to illustrate the changing P-values at sample cut-off values (Figure 2) and follow-up times (Figure 3).

A)

B)

Figure 1. P-values (height) as a function of normalised entropy cut-off and follow-up period A) The full array of P-values, B) a portion of the P-value array to give better clarity of the region in which the lowest P-values were obtained. In both plots, green signifies a P-value of < 0.05.

Figure 2. P-value as a function of normalised entropy cut-off value for the longest follow-up time (blue curve) and the follow-up time that produced the lowest P-values (22 months, red curve). P = 0.05 is indicated by the black dashed line.

Figure 3. P-value as a function of follow-up period for two normalised entropy cut-off values: blue curve: cut-off = 1.168 (the value found to produce the lowest P-values using our data), red curve: cut-off = 1.233 (the value found to be optimal by Win et al.). P = 0.05 is indicated by the black dashed line.

Figures 1 to 3 show that the P-value was found to be highly sensitive to both the choice in normalised entropy cut-off value used to dichotomise the patient as well as the length of follow-up period. Over all conditions, the lowest P-value was found to be 0.028 (obtained with a normalised entropy cut-off of 1.168 and a follow-up time of 22 months) and the maximum P-value was 0.9999 (obtained with a normalised entropy cut-off of 1.134 and a follow-up time of 5.7 months). For the longest follow-up time (which would be the time normally reported in survival studies), the lowest P-value was 0.11 and hence this is the value that would have been reported if the optimised cut-off approach were used. This would not have been considered statistically significant, even using a critical value of 0.05 (the default value often used without adjustment for use of the optimised cut-off approach or multiple hypothesis testing).

The P-value was found to be highly variable with respect to the choice in normalised entropy cut-off value. For example, at a follow-up time of 22 months, 4 results were recorded that were < 0.05 , with the lowest value ($P = 0.028$) recorded with a cut-off of 1.168 (which is therefore the value that would have been selected if the 'optimised cut-off' approach had been adopted). However if the cut-off was changed only slightly from 1.168 to 1.151, this changes the result from one that would have been considered significant ($P = 0.028$) to one that is not significant ($P = 0.173$) even though this resulted from only 3 patients moving from the low- to the high- normalised entropy group. Similar fluctuations in P were seen at other follow-up times, with no clear stable minimum with P being consistently < 0.05 .

Investigating P as a function of follow-up period also highlighted large fluctuations; the shortest follow-up times were associated with rapid changes in P-values, due to the low number of events observed and each new event consequently having a substantial impact on the outcome. Beyond around 20 months follow-up, the curve becomes smoother, although the P-value had not settled on a consistent value; a systematic rise in the P-values between approximately 30 and 60 months follow-up is seen. This is due to the survival curves initially separating but subsequently coming

closer together, so that a shorter follow-up period of 20 – 30 months yields some results that could be considered statistically significant, whereas at 40 months or above, no results were < 0.05 , irrespective of the choice of normalised entropy cut-off. This is illustrated in the Kaplan-Meier plot, shown in Figure 4 and shows that good follow-up is needed to understand the nature of differences between survival curves.

Figure 4. Kaplan-Meier plot produced using a normalised entropy cut-off of 1.233 (as used by Win et al.). The patients with a normalised entropy > 1.233 appear to have a slightly longer survival than the ≤ 1.233 group by around 20 – 30 months, but the curves have rejoined by around 50 months.

Comparison with Win et al results

The P-value obtained when the Win et al conditions were replicated (normalised entropy cut-off = 1.233, follow-up period = 40 months) was 0.162. Their result under the same conditions obtained with the validation dataset was $P < 0.001$. Increased follow-up for our data increased the P-value further to 0.257. Therefore we were unable to replicate Win et al’s results, since our P-value under equivalent conditions was much greater than they observed. These differences mean that while Win et al concluded that CT-derived tumour heterogeneity (as determined using normalised entropy) is an independent predictor of survival in addition to stage in NSCLC, our data have not demonstrated the same prognostic power. The possible reasons for these differences are outlined in the discussion section.

Simulated Data

The results obtained from patient data can be compared with simulated data in Matlab to illustrate the difficulties in drawing conclusions in scenarios where the P-value is highly dependent on the follow-up time. Figure 5 shows survival curves of 51 simulated patients (to match our real sample size), split into two groups, with the survival of the patients in both groups being sampled from a normal distribution of mean 50 months and standard deviation 20 months. Since the survival of the two groups was defined in the same way, any difference in survival seen is coincidental. In this case,

the P-value for the longest follow-up (taking into account the full Kaplan-Meier curves) was 0.55, which would not have been considered statistically significant if this result had been produced in a real study. Had the P-value been assessed at an earlier follow-up time however, a much lower P-value could have been produced, with a minimum of 0.019 occurring with a follow-up time of 50.6 months. In a similar manner to our data, the survival curves appear to separate for some time before re-joining, causing an initial decrease in P-value, followed by a steady increase. This same pattern seen in our data could therefore have been obtained by chance, with the low P-values seen at intermediate follow-up times not being indicative of a real difference between the low- and high-normalised entropy groups, but instead caused by chance.

Figure 5. Left: Kaplan-Meier curves of simulated survival data where the modelled survival of groups 1 and 2 are from an identical population. Right: The P-value as a function of follow-up period for the Kaplan-Meier plot shown on the left, with a minimum of 0.019 observed at close to 50 months, but rising to 0.55 by the end of the follow-up period.

Alternatively, it is possible for a genuine difference to be present between the two groups, but statistical significance is not reached because insufficient data are available to be certain that the differences are real. Clearly this scenario is most likely where the difference between the groups is small. This is illustrated in Figure 6, which was again simulated with survival being characterised by a normal distribution; group 1's normal distribution has a mean of 50 months and group 2 has a mean of 45 months. The standard deviation for both groups was 20 months. Therefore in this case there is a genuine difference between the groups, but like the case shown in Figure 5, the P-value obtained at the longest follow-up time (0.42) would not have been considered statistically significant had it been obtained in a real study. Also like the previous case, a minimum P-value is reached with an intermediate follow-up ($P = 0.026$ at 56.4 months), beyond which a steady increase in P is seen. This is not to say that this curve shape is expected in survival studies; these particular simulations have been selected from the set described in the methods to illustrate the difficulties in interpreting P-value results when they are sensitive to the follow-up period.

Figure 6. Left: Kaplan-Meier curves of simulated survival data where the modelled mean survival of group 1 is 5 months longer than for group 2. Right: The P-value as a function of follow-up period for the Kaplan-Meier plot shown on the left, with a minimum of 0.026 observed at 56.4 months, but rising to 0.42 by the end of the follow-up period.

Discussion

Patient data – varying input parameters

Varying the cut-off and follow-up time resulted in large variations in the P-value, often very rapidly with respect to changes in the input parameter. No clear stable minimum was obtained with consistently low P-values. Some of the P-value results would have been considered statistically significant if compared against a critical value of $P < 0.05$: altogether 39 / 2550 results met this criterion. However, these were isolated results, with other results of > 0.05 obtained from very similar conditions. Note that we are not stating that the results that were < 0.05 are statistically significant (and that the technique therefore has prognostic power under those conditions) rather that these are the conclusions that may have been reached had a study been performed where one of these P-values had been produced.

The Kaplan-Meier plot using the Win et al parameters (in Figure 4) show an initial separation of the high- and low-normalised entropy groups (producing low P-value results at intermediate follow-up times), followed by a rejoining of the two groups after further follow-up. It is not possible to ascertain from these results whether the initial gap is real (i.e. there is a genuine difference in survival between the groups) and the small number of patients was the reason for significance not being reached, or whether this was seen by chance and that the survival in the two groups is actually equivalent. A larger cohort of patients would be needed to determine which of these scenarios pertains to the data collected here. It should be noted though that the sample size of 51 in this study is comparable (Win et al 2013, Miles et al 2009, Cook et al 2013, Ganeshan et al 2012b, Ng et al 2013, Cheng et al 2013, Ravanelli et al 2013) or greater than (Willaime et al 2013, Tixier et al 2011, Yip et al 2014, Goh et al 2011, Ganeshan et al 2012a, Vadiya et al 2012) many of the studies in the

published literature. (The above studies involved between 15 and 56 patients.) Therefore the issues of false-negative findings due to sample size are also applicable to a large number of the published studies in this field, and should be taken into consideration when evaluating studies with small sample sizes.

Comparison with Win et al

There are several reasons why the differences between the findings in this study ($P > 0.05$ for a normalised entropy cut-off of 1.233 and follow-up of 40 months) and those of Win et al ($P < 0.001$ for the validation study using the same input parameters) may have occurred. It may be that the finding by Win et al was a false positive, and hence the results could not be replicated in another study. It could also be that normalised entropy has real prognostic power but that negative results were obtained in our study, either because of differences in the study designs, or because we obtained a false-negative as a result of the patient sample size. Further validation studies would be needed to address false-positive or false-negative issues, but to date there are no further validation studies pertaining to the use of normalised entropy for lung cancer patients and hence it is not possible to say definitively what the cause of the difference is.

If different results were obtained because of differences in study design, this would suggest that both results could be genuine, meaning that normalised entropy could have prognostic value in specific scenarios, but not in others. The main difference between the two studies is in the patient cohort as seen in Table 1; the Win et al study involved a large number of stage I patients (38 % of all patients) compared with our patient cohort (6 % were stage I). Conversely our patient cohort was more weighted to stage III (67 % of all patients, compared with 21 % for Win et al). Therefore if normalised entropy was predictive of survival only for early stage patients, this could cause the differences seen in the two study outcomes. Histologies were also slightly different, with the Win et al cohort limited to NSCLC, but only 2 SCLC patients were included in our study, so is unlikely to have had a substantial impact. Different scanners with different acquisition protocols may also have contributed to differences in study outcomes although it should be noted that the same type of system was used in both of these studies (GE Discovery) with similar acquisition parameters (see Methods for details). This is however an aspect of texture analysis that is yet to be fully explored.

Another difference in the study designs is the treatment type that the patients underwent; the patients in this study all underwent IMRT radical radiotherapy following the PET-CT imaging, whereas the Win et al study included patients undergoing palliative and radical treatments. Win et al stated that the mixture of treatment characteristics in their cohort indicated that 'texture heterogeneity was a survival predictor independent of treatment', which if indeed is the case, then the difference in study outcomes should not be attributable to the treatment characteristics.

If however it is the case that variations in the patient cohort or study details can cause a substantially different result, this has important implications for the wider use of texture analysis in the clinical setting. For texture analysis to be a clinically useful tool, parameters need to be robust to slight variations in acquisition details, etc. and the findings need to be applicable to a wide cohort of patients. However it is not possible to determine whether these aspects affected the results here, and further validation studies are needed to assess the general applicability and the robustness of each texture parameter.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

It is recommended that future texture analysis studies take these issues into account, to give more insight into the real prognostic power of these metrics in a wide patient cohort. Assessing the change in P-value as the cut-off and follow-up time are changed can give an indication of the stability and robustness of the result, to help ascertain whether the metric is likely to yield prognostic power in another setting. Furthermore, multi-centre studies are needed to determine whether the cut-off found to be optimal in one centre is still applicable in another, or whether a re-optimisation would be necessary in each new setting. If the latter, this could prove a serious limitation of the clinical utility of the technique. Until these questions are addressed, the widespread utility of texture metrics may not reach its full potential.

Simulated data

The P-value trends obtained when follow-up time was changed for the patient study (where P-values were initially noisy, then decreased to < 0.05 before gradually increasing to > 0.05) were successfully replicated with simulated data. In the first scenario this trend was generated despite the survival of the two groups being modelled in an identical way (i.e. there was no genuine difference present), whereas in the second scenario, the survival characteristics were genuinely different.

This illustrates the issues of both false positive and false negative results; the intermediate follow-up time in the first scenario produced $P < 0.05$, which would have been identified as statistically significant in many studies, despite the fact that the modelled survival characteristics of the two groups were identical. Conversely, the longest follow-up for the second scenario produced a P-value that would not have been considered significant, despite the different survival characteristics of the two groups. These scenarios become less likely as the sample size is increase, although as previously discussed, a sample size of 50 patients or less is commonly seen in the published literature and hence the outcomes obtained with these simulations could have also been observed in those published studies.

Conclusions

Results of survival analysis using a CT-derived heterogeneity parameter were found to be highly variable with respect to the choice in cut-off value to dichotomise the patient set. Therefore although some P-values less than 0.05 were obtained, the rapid fluctuations in P with respect to the cut-off value suggests that results are not robust and may not be reproducible over a heterogeneous, multi-centre patient cohort. It also highlights that if an ‘optimised cut-off’ approach is used (whereby the cut-off value is selected by minimising the associated P-value), false positive results are possible. Validation in this case is therefore essential in order to identify the parameters that have genuine prognostic value.

P-values were also found to be sensitive to the follow-up period, with results < 0.05 obtained at some follow-up times, but not others. This sensitivity should be taken into account when drawing conclusions from a survival study; if insufficient events have been recorded by the end of follow-up, the P-value is likely to be unstable and conclusions could be radically altered with the addition of extra events.

Significant results were not obtained when the follow-up and cut-off values used by Win et al were replicated, which could have implications for the wider applicability of normalised entropy as a prognostic tool in lung cancer; further validation studies are urgently needed to investigate this further.

Acknowledgement

This project was supported by Alliance Medical Limited and is part of the research collaboration with Royal Surrey County Hospital NHSFT and the University of Surrey.

References

- Alobaidli, S., McQuaid, S., South, C., Prakash, V., Evans, P. and Nisbet, A. (2014) "The role of texture analysis in imaging as an outcome predictor and potential tool in radiotherapy treatment planning", *The British Journal of Radiology*, Vol. 87(1042)
- Altman, D.G., Lausen, B., Sauerbrei, W. and Schumacher, M. (1994) "Dangers of using "optimal" cutpoints in the evaluation of prognostic factors" *J Natl Cancer Inst*, Vol. 86, pp. 829–835.
- Babyak, M.A. (2004) "What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models" *Psychosom Med*, Vol. 66, pp. 411–421
- Benjamini, Y and Hochberg, Y (1995) "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing" *Journal of the Royal Statistical Society Series B (Methodological)* Vol. 57, pp. 289–300
- Buvat, I., Orlhac, F., and Soussan, M. (2015) "Tumor (2015)Texture Analysis in PET: Where Do We Stand?" *J Nucl Med*, Vol. 56(11)
- Chalkidou, A., O'Doherty, M.J. and Marsden, P.K. (2015) "False Discovery Rates in PET and CT Studies with Texture Features: A Systematic Review" *PLoS ONE*, Vol. 10(5)
- Cheng, N.M., Fang, Y.H., Chang, J.T., Huang, C.G., Tsan, D.L., Ng SH, Wang, H.M., Lin, C.Y. and Yen T.C. (2013) "Textural features of pre- treatment 18F-FDG PET/CT images: prognostic significance in patients with advanced T-stage oropharyngeal squamous cell carcinoma", *J Nucl Med*, Vol. 54, pp.1703–1709
- Cook, G.J., Yipm C., Siddique, M., Goh, V., Chicklore, S., Roy, A., Marsden, P., Ahmad, S. and Landau, D. (2013) "Are pretreatment 18F-FDG PET tumor textural features in non-small cell lung cancer associated with response and survival after chemoradiotherapy?" *J Nucl Med*, Vol 54, pp.19–26
- Eliat, P., Olivie, D., Saikali, S., Carsin, B., Saint-Jalmes, H., de Certaines, J.D. (2012) "Can dynamic contrast-enhanced magnetic resonance imaging combined with texture analysis differentiate malignant glioblastoma from other glioblastoma?" *Neurol Res Int*, Vol. 2012

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Ganeshan, B., Miles, K.A., Young, R.C., Chatwin, C.R. (2009) "Texture analysis in non-contrast enhanced CT: impact of malignancy on texture in apparently disease-free areas of the liver" *Eur J Radiol*, Vol. 70, pp. 101–10

Ganeshan, B., Skogen, K., Pressney, I., Coutroubis, D. and Miles, K. (2012a) "Tumour heterogeneity in oesophageal cancer assessed by CT texture analysis: preliminary evidence of an association with tumour metabolism, stage, and survival" *Clin Radiol*, Vol. 67, pp. 157–64

Ganeshan, B., Panayiotou, E., Burnand, K., Dizdarevic, S and, Miles, K. (2012b) "Tumour heterogeneity in non-small cell lung carcinoma assessed by CT texture analysis: a potential marker of survival" *Eur Radiol*, Vol. 22, pp. 796–802

Goh, V., Ganeshan, B., Nathan, P., Juttla, J.K., Vinayan, A. and Miles KA (2011) "Assessment of response to tyro- sine kinase inhibitors in metastatic renal cell cancer: CT texture as a predictive biomarker" *Radiology*, Vol. 261, pp. 165–171

Hilsenbeck, S.G., Clark, G.M. and McGuire, W.L. (1992) "Why do so many prognostic factors fail to pan out?" *Breast cancer research and treatment*, Vol. 22, pp. 197–206

Holm, S. (1979) "A simple sequentially rejective multiple test procedure" *Scandinavian Journal of Statistics*, Vol. 6 (2), pp. 65-70.

Mattonen, S.A., Palma, D.A., Haasbeek, C.J., Senan, S. and Ward, A.D. (2014) "Early prediction of tumor recurrence based on CT texture changes after stereotactic ablative radiotherapy (SABR) for lung cancer" *Med Phys*, Vol. 41

Mayr, N.A., Huang, Z., Wang, J.Z., Lo, S.S., Fan, J.M., Grecula, J.C., Sammet, S., Sammet, C.L, Jia, G., Xhang, J., et al. (2012) "Characterizing tumor heterogeneity with functional imaging and quantifying high-risk tumor volume for early prediction of treatment outcome: cervical cancer as a model" *Int J Radiat Oncol Biol Phys*, Vol. 83, pp. 972–9

Miles, K.A., Ganeshan, B., Griffiths, M.R., Young, R.C. and Chatwin, C.R. (2009) "Colorectal cancer: texture analysis of portal phase hepatic CT images as a potential marker of survival" *Radiology*, Vol. 250, pp. 444–52

Ng, F., Ganeshan, B., Kozarski, R., Miles, K.A. and Goh, V. (2013) "Assessment of primary colorectal cancer het- erogeneity by using whole-tumor texture analysis: contrast-enhanced CT texture as a biomarker of 5- year survival" *Radiology*, Vol. 266, pp. 177–184

Peduzzi, P., Concato, J., Kemper, E., Holford, T.R. and Feinstein, A.R. (1996) "A simulation study of the number of events per variable in logistic regression analysis" *J Clin Epidemiol*, Vol. 49, pp. 1373–1379

Vasconcelos, V., Silva, J.S., Marques, L. and Barroso, J. (2010) "Statistical textural features for classification of lung emphysema in CT images: a comparative study" *Proceedings of the 5th Iberian Conference on Information Systems and Technologies (CISTI); 16–19 June 2010; Santiago de Compostela, Spain*. Piscataway, NJ; IEEE, pp. 1–5.

- Win, T., Miles, K. A., Janes, S.M., Ganeshan, B., Shastry, M., Endozo, R., Meagher, M., Shortman, R.I., Wan, S., Kayani, I., et al (2013) "Tumor heterogeneity and permeability as measured on the CT component of PET/CT predict survival in patients with non-small cell lung cancer" *Clinical Cancer Research*, Vol. 19(13), pp. 3591–3599
- Xu, Y., Sonka, M., Mclennan, G., Guo, J. and Hoffman, E.A. (2006) "MDCT-based 3-D texture classification of emphysema and early smoking related lung pathologies" *IEEE Trans Med Imaging*, Vol. 25, pp. 464–75
- Yip, C., Landau, D., Kozarski, R., Ganeshan, B., Thomas, R., Michaelidou, A. and Goh, V. (2014) "Primary esophageal cancer: heterogeneity as potential prognostic biomarker in patients treated with definitive chemotherapy and radiation therapy" *Radiology*, Vol. 270, pp.141–8.
- Yu, H., Caldwell, C., Mah, K. and Mozeg, D. (2009) "Coregistered FDG PET/CT-based textural characterization of head and neck cancer for radiation treatment planning" *IEEE Trans Med Imaging*, Vol. 28, pp. 374–83
- Zhang, H., Graham, C.M., Elci, O., Griswold, M.E., Zhang, X., Khan, M.A., Pitman, K., Caudell, J.J., Hamilton, R.D., Ganeshan, B. and Smith, A.D. (2013) "Locally advanced squamous cell carcinoma of the head and neck: CT texture and histogram analysis allow independent prediction of overall survival in patients treated with induction chemotherapy" *Radiology*, Vol. 269, pp. 801–809