Tina Memo No. 2017-006
Internal, rejected by Cancer Research, see also memo 2015-04 and 2018-004.

Mathematical Modeling of Tumor Heterogeneity Increases Statistical Power in Assessing Response to Therapy.

Paul D Tar, Neil A Thacker, James PB OConnor, et.al.

Last updated
9 / 03 / 2017

Centre for Imaging Sciences,
Medical School, University of Manchester,
Stopford Building, Oxford Road,
Manchester, M13 9PT.

Mathematical Modeling of Tumor Heterogeneity Increases Statistical Power in Assessing Response to Therapy.

Paul D Tar, Neil A Thacker, James PB OConnor. [1]

Key words: Biomarker; Heterogeneity; Imaging; MRI; Response; Tumor

ABSTRACT

Imaging demonstrates that preclinical and human tumors are heterogeneous. Phenotypic variation in the control group, which is more common in complex models such as patient-derived explants, can obscure detection of significant therapeutic effects in treated tumors. This can result in halting development of effective therapies due to limitations in experimental design, rather than due to therapeutic failure. A method to model biological variation and heterogeneity in imaging signals is described. Specifically, Linear Poisson modelling (LPM) evaluated changes in apparent diffusion co-efficient (ADC) before and 72 hours after radiotherapy, in two xenograft models of colorectal cancer. Control group data enabled construction of an LPM for each xenograft model, where complex ADC distributions were modelled in their entirety as linear combinations of multiple probability mass functions, each with their own weighting factors and sampled according to Poisson statistics. Leave-one-out analysis of the control xenografts provided methodological validation. When the same LPMs were applied to treated tumors, the LPMs detected highly significant changes in each tumor relative to the LPM components describing untreated tumor tissue. The LPM analysis were highly significant for all tumors, equating to gain in power of 16 to18 fold, compared with cohort-level summary analysis of MRI volume, mean ADC and tissue pathology biomarkers. Furthermore, LPM enabled the relative volumes of responding and non-responding tissue to be estimated for each xenograft model. Leave-one-out analysis of the treated xenografts provided quality control and identified potential outliers, raising confidence in LPM data at clinically relevant sample sizes. In summary, LPM can remove unwanted biological variation in image data due to development of intratumoral spatial heterogeneity seen with natural history growth in control tumors. This substantially increases sensitivity to treatment-induced change, thus increasing statistical power. Major Findings This article describes an analysis approach to complex imaging data that improves the efficient detection of treatment effects relative to control, in heterogeneous tumors. The method is over an order of magnitude more sensitive than standard statistical analyses. This has important implications for the 3Rs, particularly when designing complex preclinical avatar and co-clinical trial experiments where treatment effects must be detected in multiple small cohorts of heterogeneous tumors.

INTRODUCTION

Preclinical experiments and early clinical studies are essential for understanding the fundamental mechanisms driving the growth of malignant tumors and for assessing potential anti-cancer effects of new therapies [1-3]. In general, assessments are made by measuring tumor growth curves; by evaluating cell or plasma based assays or tissue pathology at one or more time points; and by non-invasive serial assessment by imaging. In all of these approaches, significance testing is performed typically on small numbers of subjects, [3,4]. Tumors are biologically heterogeneous5, . Research

[1]This draft document produced by these authours. A journal article is in preparation with the following additional authors, Isabel Peset,Muhammad Babur, Yvonne Watson, Sue Cheung, Ross A Little, Roben G Gieling, Francesca Trapani, Garry Ashton, Caron Abbey, Steve Bagley, Kaye J Williams.

studies using genomics [7], tissue pathlog [8] or clinical imaging [9] can identify and quantify spatial heterogeneity and have shown that these metrics might provide prognostic and predictive biomarkers of clinical outcome. Typically, studies measure the degree of heterogeneity within individual tumors or identify regions with certain cell populations that may mediate response to therapy and resistance[10]. However, tumor heterogeneity can also be a practical problem for studying cancer biology. In small preclinical and clinical studies, substantial spatial variation can occur in control and treatment group tumors due to natural history. This variation can obscure detection of significant biological effects of therapy, such that therapies with potential clinical benefit may be inadvertently halted in the developmental pipeline. Imaging studies generally adopt one of two approaches to derive biomarkers that measure heterogeneity within individual tumors (intratumoral heterogeneity)[9]. One approach attempts to identify the geographic subregions that drive response to therapy, subsequent resistance and relapse during treatment failure. This requires solutions to the significant challenges of both image segmentation (to identify voxels with common structural or biological features) and voxel-to-voxel registration between time points. Alternatively, imaging data can be regarded as a sample from a distribution, providing histograms where the spatial structure of a tumor is disregarded [11]. However, the complexities (e.g. non-Gaussian nature) of imaging data make it difficult to use simple histogram parameters to quantify therapy-induced changes in tumor biology [12]. Fortunately, complex data distributions of digitally encoded data (as found in imaging readouts of heterogeneous tumors) can be modeled in their entirety as linear combinations of multiple probability mass functions, each with their own weighting factors, and sampled according to Poisson statistics. Given these properties, we sought to use linear Poisson modelling (hereafter, LPM) [13] to quantify biological variation and to model uncertainties associated with data samples acquired in clinically relevant imaging methods. Specifically, we hypothesized that LPM would provide a method for assessing the volume of change within individual tumors, yielding a more powerful, efficient and sensitive method of detecting response to therapy, compared to conventional cohort-based analysis of imaging and pathology data. This benefit was anticipated since LPM not only models volumetric changes allowing estimates of the proportion of tumor changing after therapy but also removes the effect of unwanted biological variation due to tumor growth found in control data. We hypothesized that this benefit would transform the potential for image-based analyses to assess the preclinical development of novel therapeutics.

MATERIALS AND METHODS

Experiments were performed in two murine xenograft models of human colorectal cancer treated with single high dose fraction of radiotherapy (RT) or sham (control). The clinically available MRI biomarker apparent diffusion co-efficient (ADC) [14] was derived. Studies were performed in compliance with the NCRI Guidelines for the welfare and use of animals in cancer research [15] and with Licences issued under the UK Animals (Scientific Procedures) Act 1986 (PPL 40/3212) following local Ethical Committee review.

Tumor implantation

LOVO and HCT116 colorectal carcinoma cells were cultured in RPMI 1640 medium supplemented with 10% heat inactivated fetal calf serum (FCS) at 37oC in a humidified 5% CO2 incubator. Cells were passaged every 2-3 days using TEG solution (0.25% trypsin, 0.1% EDTA and 0.05% Hanks balanced salt solution in PBS). Tumor xenografts were initiated from 5 x 106 cells per mouse (in 0.1mL serum-free culture medium) injected subcutaneous in female nu/nu CBA mice aged 10 weeks old. Immediately prior to in vivo implantation, all cells tested negative for mycoplasma

infection and the number of short tandem repeats (STRs) present at 7-10 loci were assessed by PCR to provide STR profiles, from which cell line authenticity was confirmed.

Study design

Tumor size was monitored using callipers and the formula for ellipsoid volume (V), where V = (π/6/6) LWD. Here, L,WandDare the largest orthogonal dimensions of the ellipsoid. When tumors reached 300-400 mm^3 in size by calliper measurement, mice were randomised to sham or given tumor-localised RT (single 10Gy fraction) using a metal-ceramic MXR-320/36 X-ray machine (320kV, Comet AG, Switzerland). The RT was administered under ambient conditions to restrained, non-anaesthetised mice. The restrained mice were held in a lead-shielded support perpendicular to the source. Irradiation was delivered at a dose rate of 0.75 Gy/min. Mice were turned around halfway through the procedure to ensure a uniform tumor dose. Imaging was performed at baseline immediately prior to RT and 72 hours post RT along with calliper measurement of tumor volume. After the second MRI scan, animals were killed humanely by cervical dislocation, without recovery from anaesthesia.

MRI acquisition and analysis

Mice were anaesthetised with isoflurane delivered through a nose cone apparatus at 2ml/min, in 100% oxygen gas as a carrier. Respiration rate was monitored throughout the experiment by use of an electronic respiratory monitor apparatus. A heated water bed was provided to maintain the animals at constant temperature of 36°C throughout each scan. MRI was performed on a 7T Magnex instrument (Magnex Scientific Ltd, Oxfordshire, UK) interfaced to a Bruker Avance III console and gradient system (Bruker Corporation, Ettlingen, Germany), using a volume transceiver coil. Whole scan time was approximately 25 minutes per animal. Diffusion-weighted imaging (TR/TE = 2250/20ms; α = 90°b values 150, 500 and 1000 s/mm^2 along one diffusion direction; matrix 128 x 128 and FOV 2.56 x 2.56cm; 15 contiguous slices of 0.6mm thickness) was performed after localisation with a T2-weighted anatomical sequence (TR/TE = 2410/50; α = 136.8°matrix 256 x 256 and FOV 2.56 x 2.56cm; 15 contiguous slices of 0.6mm thickness). ADC maps were generated by selecting a region of interest on the lowest b value image. Voxel-wise values of ADC were calculated using in house software across the tumor using a least squares fitting routine for the equation $S = S0\ exp(-bD)$, where S0 represents the signal intensity in the absence of a diffusion sensitising gradient, S the signal intensity for a particular b value, b the numerical value in s/mm^2 and D the apparent diffusion coefficient (mm^2/s). Mean ADC was computed for all voxels across regions of interest defined in each slice, at baseline and 72 hours post sham/RT. The inter-quartile range (IQR) of voxel ADC values was also calculated on a per tumor basis, to provide a simple measurement of the heterogeneity of ADC within the tumors. To validate the ADC measurement in this protocol, measurements were verified using an ice water phantom, consisting of an inner chamber of ice water surrounded by a larger chamber of ice to maintain the inner chamber water at approximately 0°[16].

Linear Poisson modelling of ADC data

LPM is an extension of conventional pattern recognition methods (such as PCA and ICA) that build approximate models of density based upon sample data. However, LPM is implemented in the style of a conventional regression fit, enabling extracted distributions to be quantified [13]. The analysis extracts areas from significantly overlapping curves in a fit. Mathematical

modelling of distributions defines common components that together describe the overall data. The components are generated by examining the relationship between two distributions (baseline and post treatment). Although specific locations of the image data cannot be attributed with great certainty to any given curve, the LPM will nonetheless fit the quantities of each curve that best accounts for these distributions using likelihood estimation. In addition, goodness-of-fit and error estimates on computed quantities are possible through LPMs error theory, which incorporates knowledge of statistical errors in target datasets along with systematic errors due to the construction of models from training data.

Model construction and volume estimation

For control tumors, the regression was fitted to a combination of components to each ADC distribution to learn by example. The model was then extended to include variations seen in treated tumors on a case-by-case basis. Components are accordingly defined as belonging either to control data patterns or being significantly different and hence having patterns related to treatment (Box 1; Supplementary Materials and Methods). LPM was applied to each tumor on a case-by-case basis. The quantity of voxels in each class (untreated or treated components) was estimated, since the total area under the fit associated with each distribution is proportional to the quantity of tissue involved in each class. Estimates of associated errors were also derived; these are essential to enable construction of a statistical hypothesis test (i.e. that treatment has induced an effect). A chi-squared per degree of freedom (p.d.o.f.) test was also applied as a goodness-of-fit check. This should be close to unity when data is accurately described by the LPM. When the chi-squared p.d.o.f. is greater than unity, the estimated errors can be scaled [17].

Model validation

We used a combination of control testing and leave-one-out validation to provide technical validation [18] for the LPM method (Supplementary Material). Firstly, by definition there should be no significant effects of treatment in the control tumors. Treatment models were therefore fitted to control data to ensure the measured effects of treatment were consistent with zero (with error bars). Secondly, if control and treated LPM are representative of typical control or treated tumors respectively, then their application to independent data should yield equivalent results to data from which the models were original estimated. A leave-one-out analysis was performed in which multiple models were constructed, with each control tumor being excluded in turn, before being assessed as an independent sample. This leave-one-out strategy in control data enables stringent testing to be performed in numbers of data sets that are typical of those used in preclinical cancer imaging experiments [19]. This approach also provides a means to mitigate against false-positive results through quality control (i.e. representativeness testing) of training data.

Pathology analysis

Following imaging, mice were euthanized. Tumors were excised whole and bisected along the imaging plane so that the cut surface approximated to the MRI region of interest. Tumors were then fixed in 4% neutral buffered formalin for 24 hours, transferred to 70% ethanol, processed and then embedded in paraffin. Sections 5m thick were cut, floated out on a water bath, collected on charged slides and then dried at 37°overnight. Sections were stained with hematoxylin and eosin (H&E) to allow identification of viable and necrotic tumor and whole field images were obtained. Immunohistochemistry (IHC) was performed to detect: (1) apoptotic fraction, by staining for

cleaved caspase-3 (CCas3) using a 1:200 dilution CCas3 primary antibody (9661 @ 0.6ug/ml; Cell Signaling Technology, Danvers, MA); (2) total vessel density, by staining for endothelial cells using rabbit anti-mouse CD31 (Abcam ab12443 @ 4ug/ml); and (3) tumor proliferation, by staining for ki67 (Bethyl IHC00375 @ 1ug/ml. CCas3 & CD31 was detected using the refine detection kit (Leica as per manufacturers instructions) with ER1, 20 mins, 95°epitope retrieval while Ki67 was detected using Dakos Envision labelled polymer, both with citrate buffer epitope retrieval using a Biocare Pascal Decloaker (125°1 min, 90°10 secs). 3, 3-diaminobenzidine was used as the chromogen for all antibodies. All section scanning was performed using a Leica SCN400 slide scanner microscope (Leica Microsystems, Milton Keynes, UK) at 40X magnification. Pathology image analysis was performed using Definiens Developer XD version 2.5 and the Tissue Studio Portal version 4.2 (Definiens AG, Munich, Germany). Tumors were segmented into viable and necrotic tumor in each H&E or IHC stain. Percentage section area of necrosis was calculated using H&E staining. Apoptotic nuclei were detected and classified as positive or negative based on user-defined CCas3 staining thresholds per model, allowing for calculation of percentage apoptosis. Vessel density was calculated by the percentage of area viable tumor stained positive for CD31; in addition tumor cell density was determined on the counter stain. Proliferating cells were detected by Ki67 staining, with nuclei detected by counter staining in the viable tumor and classified as positive of negative based on user-defined Ki67 staining thresholds per model. Summary of the pathology analysis method is shown in Supplementary Figure 1.

Statistical analysis

Baseline and change in tumor volume and ADC (mean value and IQR) parameters were compared between control and treated tumors using Students t test for independent samples in IBM SPSS Statistics v.22 (Armonk, NY). Similar tests were used for ex vivo pathology analyses. All tests were two tailed. These tests were performed and combined to provide comparison with the statistics derived from LPM. In all tests, p¡0.05 was considered to indicate statistical significance. Corrections for multiple comparisons were applied where necessary.

RESULTS

Quantitative ADC measurements are accurate and precise

Mean ADC values measured in iced water phantom over five experiments were 1.13 x10-3 mm2/s (0.023 x10-3 mm2/s) in agreement with previous data 20. The test-retest of ADC values was determined by scanning nine HCT116 xenograft tumors twice in rapid succession using the same scan session and same operator. The wCV was 0.092%. These data show that the ADC measurements derived in the study are sufficiently accurate and precise for use in the subsequent analysis.

Cohort volumetrics, summary ADC and tumor cell density detect RT response

Conventional preclinical experiments that test therapeutic response typically use growth characteristics, imaging and pathology analyses to detect significant effects of treatment. We designed an experiment that was expected to induce a significant treatment effect that could be detected by calliper measurement, MRI volumetrics and quantitative ADC and by pathology assays. Calliper

measurement and MRI demonstrated significant growth inhibition was induced by RT in both xenograft models at 72 hours (both p=0.001), relative to control. In the LOVO xenografts, RT increased mean ADC value (p=0.0004) and increased IQR of the ADC distribution (p=0.041), relative to control. Likewise, in the HCT116 xenografts, RT increased mean ADC value (p=0.0009) and increased IQR of the ADC distribution (p=0.047), relative to control (Figure 1A). Pathology analysis showed highly significant RT-induced reductions in the density of tumor cells in both xenograft models (Figure 1B). The LOVO cohort (control 8.30 x103 nuclei/m2 versus RT 7.01 x103 nuclei/m2; p=0.004) and HCT116 cohort (control 6.41 x103 nuclei/m2 versus RT 5.69 x103 nuclei/m2; p=0.0003) showed similar extent of change. In the HCT116 cohort, significant increases were also observed in necrosis (control 32.1% versus RT 45.5%; p=0.013) and apoptosis (control 1.09% versus RT 1.86%; p=0.018). Similar differences in necrosis and apoptosis were seen in the LOVO cohorts but did not reach statistical significance. Proliferation and vessel density did not differ in the control and therapy cohorts in either xenograft model. Collectively, these data show that the significant changes were induced by RT within 72 hours in subcutaneous LOVO and HCT116 xenograft tumor models. Treatment effects were detected at the cohort level by MRI volumetrics, by quantitative ADC and by pathology assays, particularly change in tumor cell density. Effects were demonstrated to have statistical significance in cohorts of small numbers of animals, typically used in preclinical experiments.

LPM identifies the varying complexity of different xenograft models

The appropriateness of using Poisson statistics to describe the sampling of the ADC distribution was confirmed by analysing Bland-Altman (funnel) plots (Supplementary Figure 2). Optimum binning for histograms describing the ADC distributions were determined as having 200 bins (Supplementary Figure 3). This binning was used for the remainder of the study. For each xenograft model, an LPM was constructed independently and the number of model components was selected on the basis of leave-one-out cross validation. This yielded 3 components to describe control ADC distributions in the LOVO control tumors (denoted C1Lovo, C2Lovo and C3Lovo). An equivalent and independent process was performed for the HCT116 tumors. This yielded 4 components in control tumors (denoted C1HCT, C2HCT, C3HCT, C4HCT) (Figure 2). The LPM data indicates that the HCT116 xenografts were more spatially complex than the LOVO xenografts and that LPM can detect this differing level of tumor complexity. Visual inspection of ADC maps corroborated this assertion, and largely reflected a high value ADC component present only in the HCT116 tumors.

LPM technical validation identifies outliers in control groups

We used a leave-one-out approach to validate the ability of the model to distinguish data with a different ADC distribution, characterized by new components. To do this, we applied fully trained models (i.e. without excluding any data points) to both LOVO and HCT116 control xenograft data. In cohorts of around 10, all control tumors would be expected to have Z scores of less than 2. This was found for all but two tumors, with average Z scores from full models of 1.05 for LOVO and 0.94 for HCT116 (Tables 1 and 2). Differences between alternative models (leave-all-in and each of the various leave-one-out possibilities) were statistically equivalent, implying that estimated volumes were the same, within limits of estimated errors (Supplementary Tables 1 and 2). These data show that the model performs as expected, correctly accounting for each control tumor distribution as being constructed of components from untreated voxel values. The leave-one-out approach not only validates the LPM method, but also identifies outlier data in the control cohort. LOVO control tumor 4 showed a Z score of 2.94 for estimated treatment volume

and HCT116 control tumor 12 showed a Z score of 2.04, implying significant difference from other control data. This could be explained by the data being an atypical, yet otherwise valid, control sample, which could have been better modelled using additional training data. For the current study, we elected to leave these data in the control group, to impose a worst case scenario on our data, since we are describing a new methodology. More reasonably, this can be explained by these two control tumors being outliers. Therefore, LPM with leave-one-out validation enables statistically robust identification of outliers in control data, which can be a critical step in avoiding equivocal results in small low-powered preclinical studies.

LPM quantifies the percentage responding volume in each tumor

The two LPMs were extended to tumors treated with RT. For LOVO the model selection revealed 2 further components (denoted C4Lovo and C5Lovo). In HCT116 xenografts treated with RT, a further 5 components were identified (denoted C5HCT, C6HCT, C7HCT, C8HCT, C9HCT). These components represent changes in ADC distributions which were significantly different from control model ADC variations. The minimum points of chi-squared p.d.o.f. were equivalent within error bars (all approximately 10 chi-squared p.d.o.f.; Supplementary Figure 4) to those seen for control tumors, indicating comparable accuracy of LPM in all settings. Non-responding tumor was defined by the sum of the control model component volumes and responding tumor was defined by the sum of the treated model component volumes. The proportion of tumor changing with therapy was calculated, along with error bars (Figure 3). All LOVO and HCT116 tumors treated with RT showed statistically significant volumes of responding tumor. For LOVO, proportion of volume responding to RT varied between 27.6 to 68.6% (median responding volume 40.4%). For HCT116, proportion of volume responding to RT varied between 22.7 to 84.4% (median responding volume 61.4%). In comparison, all control tumors (except outlier LOVO control tumor 4) had responding volumes consistent with zero (Figure 3).

LPM biomarkers of response are more powerful than conventional analyses

LPM was investigated as a method of improving sensitivity of detecting RT-induced changes in the LOVO and HCT116 xenograft models of cancer. In LPM, the error estimates incorporate systematic processes associated with learning the model parameters (i.e. linear components), as well as the statistical errors on weighting factors used to describe each case (see Supplementary Material). LPM can capture the uncertainties on the distribution components and the weighting factors using the error estimates provided by the method. This enables construction of hypothesis tests for individual data sets, by testing the null hypothesis on a case by case basis. The probability of the treatment volume (QTREATED) being consistent with zero on the basis of estimated error was measured. The LPM approach implicitly combines information from volume and ADC change. To ensure a fair comparison between LPM and conventional measures, we combined the significance figures for conventional volume and ADC (mean and IQR), giving a total Z score of 5.2 standard deviations for LOVO (Table 3). LPM results showed higher Z score and more significant p values for many of the individual treated tumors compared to the conventional cohort-level statistics for imaging biomarkers. The combined Z score from the LPM was 21.8 standard deviations for LOVO tumors treated with RT. Since a linear increase in Z score requires a quadratic increase in data quantity, approximately 17-18 times more data (square of 21.8/5.2) would be needed for LOVO tumors to demonstrate the same treatment effect with equivalent power using volume and mean ADC compared to LPM. An equivalent comparison of summary statistics and LPM statistics in HCT116 xenografts treated with RT showed a similar gain in

statistical power, of approximately 16 fold (square of 32.6/8.1; see Table 4). These data reveal that mathematical modelling of imaging data through LPM enables substantial increase in statistical power to detect response to therapy.

DISCUSSION

In this study we describe how modelling the spatial heterogeneity present in imaging data can increase statistical power of identifying response to therapy. We investigated a technique called linear Poisson modelling in a well understood biological paradigm. Several studies have reported acute increase in the imaging biomarker ADC and corresponding change in histopathology biomarkers following RT [21-24]. We demonstrated similar changes in two xenograft models in our experiment. Importantly, imaging and pathology biomarkers varied in their ability to detect these changes. Pathology assessment of necrosis and apoptosis revealed similar magnitude differences between control and treated groups in the two xenografts models, but statistical significance was only found in a larger sample size (HCT116 tumors). In distinction, tumor cell density changes and changes in tumor size and mean ADC were so marked that both LOVO (n=8 versus 10) and HCT116 (n=13 versus 15) showed significant changes. Next, we demonstrated that LPM could appropriately describe ADC distributions of varying complexity, across two untreated xenograft models. An optimum binning was selected and model performance was assessed. We then showed three important advantages of applying LPM to analyze the ADC data, all of which would not be possible using conventional image analysis methods. Firstly, in providing method technical validation, through a leave-one-out approach, we showed that it was possible to detect outliers in control groups. It is common to have variation in control group imaging biomarker values and this can substantially limit the ability of any biomarker to detect biological differences between small cohorts of control and treated animals [25]. In the era of personalized medicine that employs tumor models of increasing biological relevance and complexity 26, the ability to exclude atypical tumors from cohort-wise analysis is of increasing importance. LPM enables outliers to be identified and excluded based on robust statistical methods. Secondly, any pair (pre- and post-) of ADC values can be assigned a probability that they are associated with variation observed within the control group, or are statistically different and thus can be considered belonging to a treatment group. By calculating the volume of voxels in each category, LPM quantifies the minimal amount of responding tissue (i.e. a lower bound) that can be detected; more voxels may respond, but cannot be distinguished from non-responding voxels within the distribution overlapping with control. Here all tumors showed some response, but the range of the lower bound on responding volumes varied by approximately 2.5 fold in LOVO and approximately four-fold in HCT116. Thirdly, this feature enables response detection on a sample by sample basis, without the need for spatial mapping. This is possible since LPM models variation within control data and then can account for this in the treatment group, identifying which voxels are different. The key finding of this study was that LPM is substantially more powerful than conventional cohort-based statistical methods for analysing imaging data. Indeed, approximately 16-18 times as much data from conventional analyses (size and mean ADC) would be required to detect changes with equivalent power compared to an LPM analysis. The implications of these data are substantial. Once a control cohort model is established, the need for similar animal numbers in the treatment group is diminished considerably. Subsequent studies for a known animal model would require a small number of new control animals (to establish equivalence with banked control data). Then very small cohorts can be tested for a given therapy. In particular, LPM can identify response on a per tumor basis with greater significance than seen in a conventional t-test analysis of control versus treatment cohorts. This simultaneously would allow reduction in animal numbers with welfare benefits[15], along with ability to identify individual responders in small studies of therapies where different tumors with varying biology are treated. This may be attractive for avatar studies where patient

derived samples are used to generate PDx and CDx models [27] and in co-clinical trials where multiple therapies are tested against animal models with different genetic knockdown/knockout featres [3]. The LPM method described here has some limitations. In its current form, LPM does not determine response to therapy at the voxel level with sufficient statistical certainty to enable generation of response maps. Data presented show clearly that mapping is not required to demonstrate with great statistical power that therapy-induced change has occurred. Given voxel correspondences between parameter distributions (with a pre- to post-treatment non-rigid registration), it would be possible to use LPM to map probability of where change has occurred. However, more signals would reduce ambiguity (i.e. increase the certainty that a given voxel has changed or has not changed) and further work is required to achieve this. As the volume of responding tissue is computed by excluding all variation which cannot be interpreted as normal control development, this value is strictly a lower bound. This bound however, is appropriate for use as part of the null hypothesis test. Our method determines this estimate without labelling individual voxels of data, but instead operates by fitting the entire data ADC distributions, learning the correlations between those from two time points. In so doing LPM can estimate the volume of treatment response without having to solve the ill-posed problem of voxel to voxel registration where investigators attempt to produce one-to-one mapping between voxels from images at different time points in tumors that change in shape and volume over time [12]. This does however prevent LPM in its current form generating voxel level treatment response maps, which might otherwise be assumed possible for a method which estimates volume of treatment response. If the control cohort is not sufficient to describe control variation then treatment volume can be overestimated by inappropriately attributing previously unseen control variation to treatment. This is the same problem as missing high sources of control variation when applying a conventional t-test, but with the problem multiplied for a higher dimensional model. Translation of the technique requires further technical and biological validation, though showing consistency in results across multiple models and therapies, with data from different laboratories [16]. Clinical application may also be possible, with collection of the necessary data in an appropriate control group. The method is protected from model construction problems that avoid over-interpretation of results. For instance, a highly atypical example will have a correspondingly high chi-squared p.d.o.f., and since quantity errors are scaled by the chi-squared p.d.o.f., the statistical significance of treatment estimates is penalized for poorly modelled data. Large quantity errors can generally be attributed to poor models, for example with few control data sets, but this problem can be reduced by adding additional (valid) training data. Equally, if contamination in the form of outliers is included in control data, the additional variability introduced in the control model reduces the ability to measure treatment, again penalising the statistical significance of results. While this reduces the statistical power of the method, it increases robustness by providing a working analysis which gives a valid, yet more limited, lower bound on volume changes. In conclusion, we have shown that LPM can remove unwanted biological variation in image data (from growth) for tumors of varying spatial heterogeneity. This substantially increases sensitivity to treatment-induced change, thus increasing statistical power. Once control models are constructed, LPM enables significant changes to be detected for single tumors. This has important implications for 3Rs (specifically reduction in animals) and LPM may facilitate design of complex preclinical avatar and co-clinical trial experiments by providing adequate power to small cohort sizes.

REFERENCES

1. Gibbs JB. Mechanism-based target identification and drug discovery in cancer research. Science 2000; 287:1969-1973

2. Conway JR, Carragher NO, Timpson P. Developments in preclinical cancer imaging: innovating

the discovery of therapeutics. Nat Rev Cancer 2014; 14:314-328.

3. Clohessy JG, Pandolfi PP. Mouse hospital and co-clinical trial project–from bench to bedside. Nat Rev Clin Oncol 2015; 12:491-498.

4. Workman P, Aboagye EO, Chung YL, Griffiths JR, Hart R, Leach MO, et al. Minimally invasive pharmacokinetic and pharmacodynamic technologies in hypothesis-testing clinical trials of innovative therapies. J Natl Cancer Inst 2006; 98:580-598.

5. Heppner GH. Tumor heterogeneity. Cancer Res 1984; 44:2259-2265.

6. Bedard PL, Hansen AR, Ratain MJ, Siu LL. Tumour heterogeneity in the clinic. Nature 2013; 501:355-364.

7. Alizadeh AA, Aranda V, Bardelli A, Blanpain C, Bock C, Borowski C, et al. Toward understanding and exploiting tumor heterogeneity. Nat Med 2015; 21:846-853.

8. Gurcan MN, Boucheron LE, Can A, Madabhushi A, Rajpoot NM, Yener B. Histopathological image analysis: a review. IEEE Rev Biomed Eng 2009; 2:147-171.

9. O'Connor JP, Rose CJ, Waterton JC, Carano RA, Parker GJ, Jackson A. Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome. Clin Cancer Res 2015; 21:249-257.

10. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med 2012; 366:883-892.

11. Just N. Improving tumour heterogeneity MRI assessment with histograms. Br J Cancer 2014; 111:2205-2213.

12. O'Connor JPB. Cancer heterogeneity and imaging. Semin Cell Biol Devel 2016

:http://dx.doi.org/10.1016/j.semcdb.2016.1010.1001.

13. McCullagh P, Nelder JA. Generalized Linear Models. London: Chapman & Hall, 1989.

14. Padhani AR, Liu G, Koh DM, Chenevert TL, Thoeny HC, Takahara T, et al. Diffusion-weighted magnetic resonance imaging as a cancer biomarker: consensus and recommendations. Neoplasia 2009; 11:102-125.

15. Workman P, Aboagye EO, Balkwill F, Balmain A, Bruder G, Chaplin DJ, et al. Guidelines for the welfare and use of animals in cancer research. Br J Cancer 2010; 102:1555-1577.

16. Doblas S, Almeida GS, Ble FX, Garteiser P, Hoff BA, McIntyre DJ, et al. Apparent diffusion coefficient is highly reproducible on preclinical imaging systems: Evidence from a seven-center multivendor study. J Magn Reson Imaging 2015; 42:1759-1764.

17. Tar PD, Thacker NA. Linear Poisson Models: A Pattern Recognition Solution to the Histogram Composition Problem. Annals of the BMVA 2014; 1:1-22.

18. OConnor JPB, Aboagye EO, Adams JE, Aerts HJWL, Barrington SF, Beer AJ, et al. Imaging Biomarker Roadmap for Cancer Studies Nat Rev Clin Oncol 2016; 14:169-186.

19. Bernsen MR, Kooiman K, Segbers M, van Leeuwen FW, de Jong M. Biomarkers in preclinical cancer imaging. Eur J Nucl Med Mol Imaging 2015; 42:579-596.

20. Chenevert TL, Galban CJ, Ivancevic MK, Rohrer SE, Londy FJ, Kwee TC, et al. Diffusion coefficient measurement using a temperature-controlled fluid for quality control in multicenter studies. J Magn Reson Imaging 2011; 34:983-987.

21. Babsky AM, Hekmatyar SK, Zhang H, Solomon JL, Bansal N. Application of 23Na MRI to monitor chemotherapeutic response in RIF-1 tumors. Neoplasia 2005; 7:658-666.

22. Henning EC, Azuma C, Sotak CH, Helmer KG. Multispectral tissue characterization in a

RIF-1 tumor model: monitoring the ADC and T2 responses to single-dose radiotherapy. Part II. Magn Reson Med 2007; 57:513-519.

23. Larocque MP, Syme A, Yahya A, Wachowicz K, Allalunis-Turner J, Fallone BG. Monitoring T2 and ADC at 9.4 T following fractionated external beam radiation therapy in a mouse model. Phys Med Biol 2010; 55:1381-1393.

24. Chung C, Jalali S, Foltz W, Burrell K, Wildgoose P, Lindsay P, et al. Imaging biomarker dynamics in an intracranial murine glioma study of radiation and antiangiogenic therapy. Int J Radiat Oncol Biol Phys 2013; 85:805-812.

25. de Jong M, Essers J, van Weerden WM. Imaging preclinical tumour models: improving translational power. Nat Rev Cancer 2014; 14:481-493.

26. Sharpless NE, Depinho RA. The mighty mouse: genetically engineered mouse models in cancer drug development. Nat Rev Drug Discov 2006; 5:741-754.

27. Malaney P, Nicosia SV, Dave V. One mouse, one patient paradigm: New avatars of personalized cancer therapy. Cancer Lett 2014; 344:1-12.

FIGURE 1

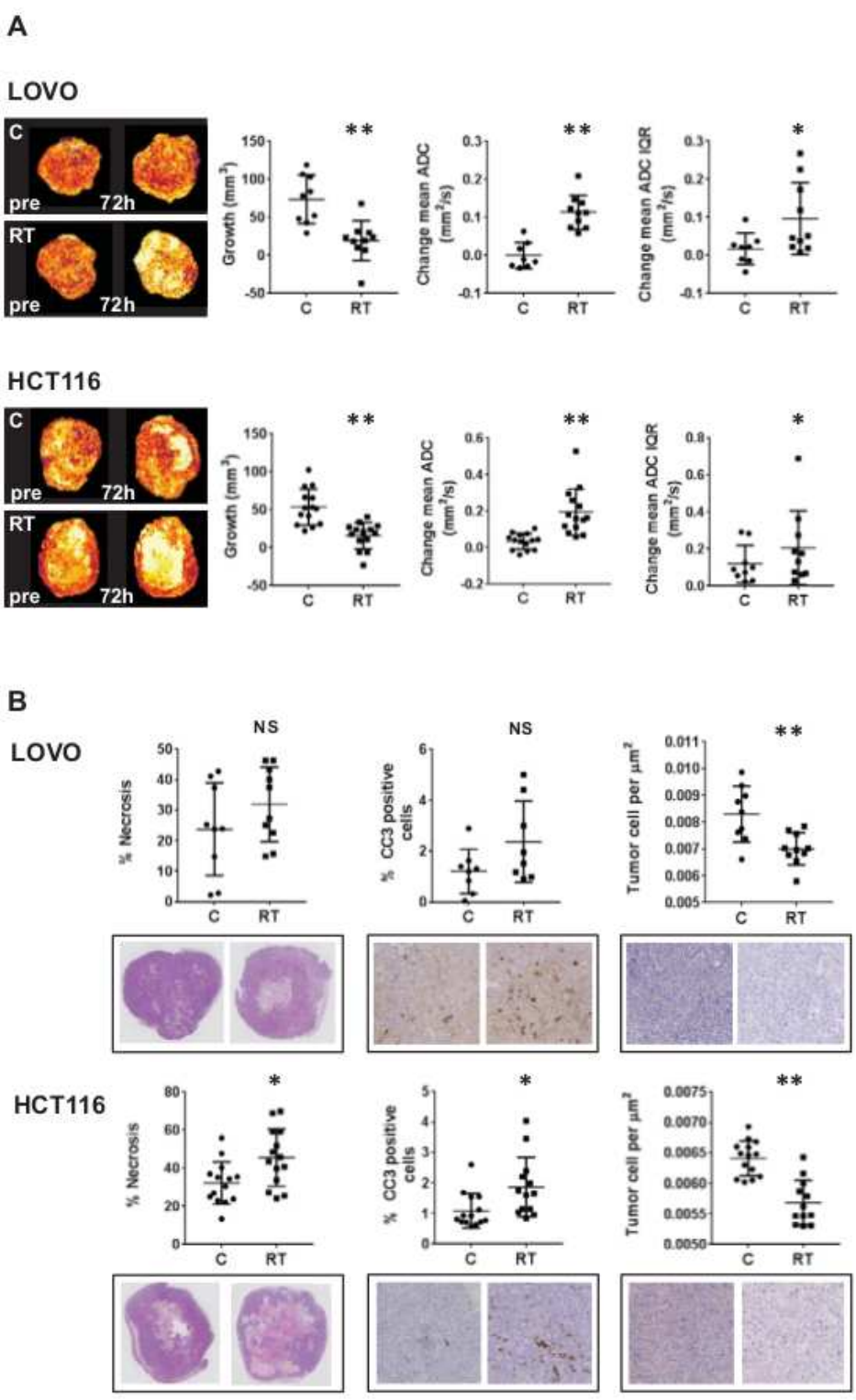

Figure 1: Figure 1: Conventional cohort volumetrics, summary ADC and tumor cell density detect RT response Statistically significant changes are detected following RT in LOVO and HCT116 tumors. A, Tumor volume, mean ADC and inter-quartile range of the ADC distribution, and in B, Tumor cell density detected on CD31 counterstain. Changes in percentage necrosis and percentage apoptosis were significant only for HCT116 tumors.

FIGURE 2

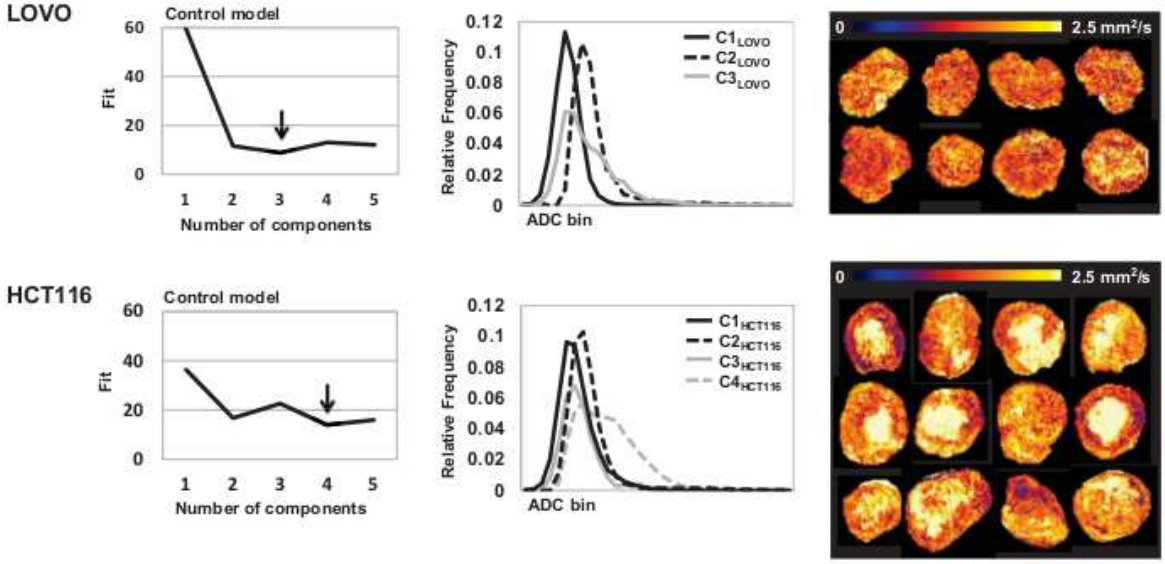

Figure 2: Figure 2: LPM identifies the varying complexity of different xenograft models The LPM is constructed in each xenograft model, based on control tumor data. Three components describe the LOVO data ADC distribution, whereas four components are required to describe the HCT116 data ADC distribution. Left hand panel: the y-axis shows the variance on residuals averaged over all leave-one-out combinations. Arrows highlight the optimum number of components for each model. Middle panel: Plots of the components by ADC value and relative frequency. Right hand panel: ADC maps show that LOVO tumors appear relatively homogeneous compared to the more spatially heterogeneous HCT116 tumors, with an additional component reflecting high values of ADC.

FIGURE 3

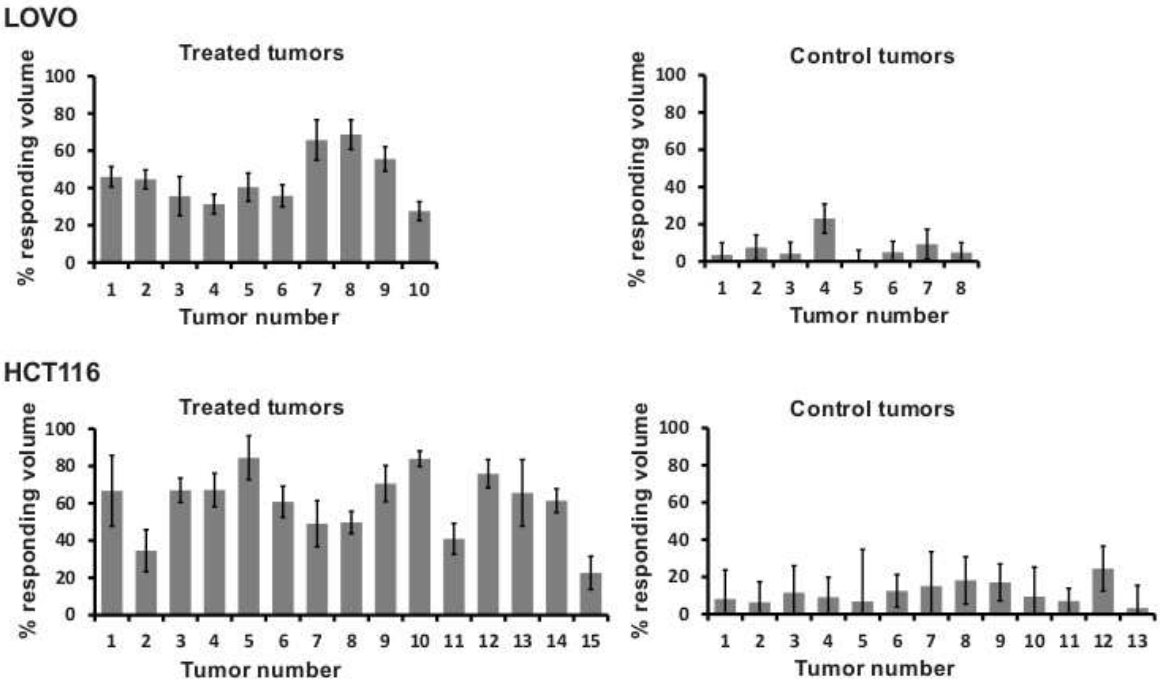

Figure 3: Figure 3: LPM quantifies the percentage responding volume in each tumor The percentage of tumor responding to therapy is calculated for each tumor (with error bars) for both LOVO and HCT116 tumors. For comparison, control group data is shown.

TABLES

Table 1: Z scores and p values for control group LOVO tumors

Mouse ID	Z score (full model)	P value	Z score (leave-one-out)	P value
1	0.54	0.588	1.27	0.205
2	1.12	0.264	1.02	0.309
3	0.68	0.499	0.55	0.579
4	2.94	0.003	5.20	<0.001
5	0.19	0.847	0.03	0.974
6	0.84	0.401	0.35	0.724
7	1.17	0.243	0.40	0.687
8	0.89	0.371	0.58	0.561

Table 2: Z scores and p values for control group HCT116 tumors

Mouse ID	Z score (full model)	P value	Z score (leave-one-out)	P value
1	0.53	0.598	0.17	0.863
2	0.58	0.559	0.38	0.704
3	0.80	0.424	1.06	0.291
4	0.85	0.393	0.76	0.449
5	0.23	0.807	0.37	0.713
6	1.44	0.151	1.90	0.058
7	0.82	0.411	0.57	0.567
8	1.44	0.149	2.25	0.025
9	1.73	0.084	1.40	0.160
10	0.60	0.551	0.88	0.377
11	1.02	0.307	0.84	0.399
12	2.04	0.041	3.57	<0.001
13	0.28	0.783	0.04	0.967

Table 3: Z scores and p values for LOVO tumors treated with RT

	Z score	P value
Cohort statistics		
Volume	3.3 (t 4.27)	0.001
Mean ADC	3.5 (t 5.91)	0.0004
IQR ADC	2.0 (t 2.23)	0.041
Combined volume and	5.2	0.000037

mean ADC		
Per tumor statistics		
Mouse ID		
1	8.52	<0.000001
2	8.75	<0.000001
3	3.40	0.000668
4	5.98	<0.000001
5	5.34	<0.000001
6	6.09	<0.000001
7	6.08	<0.000001
8	8.65	<0.000001
9	8.54	<0.000001
10	5.41	<0.000001
Combined	21.82	<0.000001

Table 4: Z scores and p values for HCT116 tumors treated with RT

	Z score	**P value**
Cohort statistics		
Volume	4.65	0.0008
Mean ADC	4.33	0.0009
IQR ADC	2.09	0.047
Combined volume and mean ADC	8.1	0.000056
Per tumor statistics		
Mouse ID		
1	3.51	0.000454
2	3.06	0.002239
3	10.26	<0.000001
4	7.48	<0.000001
5	7.16	<0.000001
6	7.28	<0.000001
7	3.97	0.000072
8	8.28	<0.000001
9	7.30	<0.000001
10	20.06	<0.000001
11	4.92	<0.000001
12	9.96	<0.000001
13	3.67	0.000243
14	9.66	<0.000001
15	2.54	0.011147
Combined	32.62	<0.000001