Bioinformatics, 34(15), 2018, 2625–2633
doi: 10.1093/bioinformatics/bty115
Advance Access Publication Date: 14 March 2018
Original Paper

OXFORD

Data and text mining

A new method for the high-precision assessment of tumor changes in response to treatment

P. D. Tar[1,*], N. A. Thacker[1], M. Babur[2], Y. Watson[1], S. Cheung[1], R. A. Little[1], R. G. Gieling[2], K. J. Williams[2,3] and J. P. B. O'Connor[3,4]

[1]Division of Informatics, Imaging and Data Science, [2]Division of Pharmacy and Optometry, Manchester Pharmacy School, Manchester M13 9PT, UK, [3]Division of Cancer Sciences, University of Manchester and [4]Department of Radiology, The Christie Hospital NHS Trust, Manchester M20 4BX, UK

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on July 27, 2017; revised on February 5, 2018; editorial decision on February 25, 2018; accepted on March 12, 2018

Abstract

Motivation: Imaging demonstrates that preclinical and human tumors are heterogeneous, i.e. a single tumor can exhibit multiple regions that behave differently during both development and also in response to treatment. The large variations observed in control group, tumors can obscure detection of significant therapeutic effects due to the ambiguity in attributing causes of change. This can hinder development of effective therapies due to limitations in experimental design rather than due to therapeutic failure. An improved method to model biological variation and heterogeneity in imaging signals is described. Specifically, linear Poisson modeling (LPM) evaluates changes in apparent diffusion co-efficient between baseline and 72 h after radiotherapy, in two xenograft models of colorectal cancer. The statistical significance of measured changes is compared to those attainable using a conventional *t*-test analysis on basic apparent diffusion co-efficient distribution parameters.

Results: When LPMs were applied to treated tumors, the LPMs detected highly significant changes. The analyses were significant *for all tumors*, equating to a gain in power of 4-fold (i.e. equivalent to having a sample size 16 times larger), compared with the conventional approach. In contrast, highly significant changes are only detected at a cohort level using *t*-tests, restricting their potential use within personalized medicine and increasing the number of animals required during testing. 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.

Availability and implementation: TINA Vision open source software is available from www.tina-vision.net.

Contact: paul.tar@manchester.ac.uk

Supplementary information: Supplementary data are available at Bioinformatics online.

1 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 (Clohessy and Pandolfi, 2015; Conway *et al.*, 2014; Gibbs, 2000). 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 (Clohessy and Pandolfi, 2015; Workman *et al.*, 2006). This can result in low statistical power. This is especially true in cases where data are complex and variable. The limitations of small sample sizes motivate the need for efficient use of data.

Differences between groups (i.e. control versus treatment, or one thearapy versus another) are most commonly assessed using *t*-tests, analysis of variance (ANOVA) or correlation analyses, which have long been available within statistical packages (Kibby, 1986). These statistical approaches are used as standard within clinical and preclinical work because they facilitate estimation of confidence intervals, Z-scores and *P*-values. These are essential outputs for the assessment of treatment responses. However, these methods assume Gaussian distributed data, which can be impossible to corroborate using the small sample sizes often used within studies. More sophisticated modern pattern recognition approaches are experimentally applied to biomedical data for the purposes of prediction (e.g. Xia *et al.*, 2017), image segmentation (e.g. Zeng *et al.*, 2017) and data mining (e.g. Zong *et al.*, 2017), for example. However, they do not generally provide conventional confidence assessments (e.g. *P*-values) and are therefore restricted to preliminary proof-of-concept rather than clinical use.

Tumors are biologically heterogeneous, leading to numerous modes of data variability that complicate analysis (Bedard *et al.*, 2013; Heppner, 1984). Research studies using genomics (Alizadeh *et al.*, 2015), tissue pathology (Gurcan *et al.*, 2009) or clinical imaging (O'Connor *et al.*, 2015) identify and quantify spatial heterogeneity and have shown that heterogeneity 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 (Gerlinger *et al.*, 2012). 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. 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. To mitigate against this, information must be accumulated over larger sample sizes to boost statistical power or unwanted sources of variability must be modeled.

Imaging studies generally adopt one of two approaches: one approach attempts to identify the geographic sub-regions 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. Pattern recognition techniques are often applied to solve these problems. In the presence of heterogenous control variability, it could be argued that an adversarial deep learning approach could be applied to classify treatment-affected voxels by learning the invariant characteristics of treatment, while ignoring the confounding changes of normal development. However, a segmented image derived through such an approach is not necessarily the best starting point for determining overall treated volumes and *P*-values for the significance of changes. Additionally, training a deep learning system typically requires far more data than is available in our problem domain.

In a second approach, imaging data can be regarded as a sample from a distribution, providing histograms where the spatial structure of a tumor is disregarded (Just, 2014). 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 (O'connor, 2017). In this approach, basic distribution parameters, such as normalization (e.g. volume of tumor), mean values or the location of percentiles are often used in conjunction with *t*-tests, ANOVA, and so on. This results in a large amount of information being discarded regarding the exact shape and behavior of individual histograms.

Given the properties of our target data (histograms of Poisson samples), we sought to use linear Poisson modeling (hereafter, LPM) (Deepaisarn *et al.*, 2017; Tar and Thacker, 2014; Tar *et al.*, 2015) to quantify biological variation and to model uncertainties associated with data samples acquired in clinically relevant imaging methods. LPMs can be considered as an extension to Gaussian mixture modeling, Dempster (1977), where the Gaussian sub-distributions are replaced with arbitrary non-parametric probability functions. The probability mass functions (PMFs) required to approximate the data are determined using an independent component analysis (Comon, 1994), designed for Poisson samples. LPM is a pattern recognition method specifically for quantitative work, facilitating the estimation of confidence, Z-scores and *P*-values.

We hypothesized that LPM would provide a method for assessing the volume of change within individual tumors, yielding a more efficient and sensitive method of detecting response to therapy, compared to conventional cohort-based analysis of imaging data. This benefit was anticipated since LPM cannot only model volumetric changes—allowing estimates of the proportion of tumor changing after therapy—but can also model the effect of unwanted biological variation due to tumor growth and heterogeneity 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.

2 Materials and methods

Imaging data were acquiredfor two murine xenograft models of human colorectal cancer (LoVo and HCT116) treated with either a single high-dose fraction of radiotherapy (RT) or sham (control). About 8 and 13 controls were used for LoVo and HCT116, respectively. A futher 10 LoVo and 15 HCT116 treated tumors were imaged. These sample sizes are typical of those found within small preclinical trials. The MRI biomarker apparent diffusion co-efficient (ADC) (Padhani *et al.*, 2009) was derived for images at baseline and 72 h after RT or sham. The tumor regions within each image were manually segmented by a clinical expert (coauthor JPB O'Connor), and the distribution of ADC values within each tumor was sampled into 2 D histograms, with one axis being ADC and the other being time ($t = 0$ h and $t = 72$ h), thus recording the ADC distributions pre- and post-treatment. Figure 1 shows example spatial distributions of such tumor data before histogramming.

As a benchmark for comparison to our method, a conventional analysis was performed. Tumors were paired between baseline and $t = 72$ h, with changes in volume, in mean ADC value and in interquartile ranges computed. The per-tumor changes measured within the control groups were compared to those measured within the treatment groups using *t*-tests. Additionally, LPM was used to construct a linear model of variability in the control cohorts, which were then extended to include additional variability found within the treatment cohorts. The fully trained models were fitted to both the control and treatment groups to estimate the relative volumes

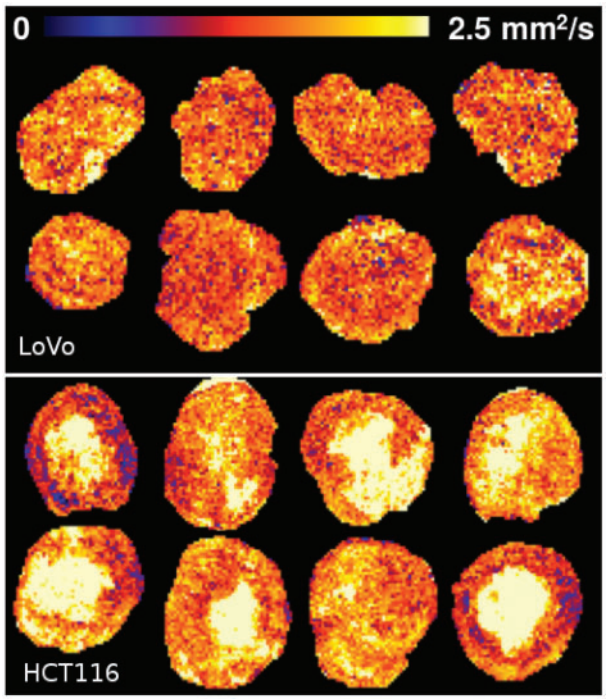

Fig. 1. Example spatial distributions of ADC values in selected tumors. Visually, HCT116 tumors are more complex and variable than LoVo tumors

associated with normal untreated tumor development and volumes associated with treatment effects. Per-tumor significances were computed, as well as cohort level significances, for comparison to the conventional *t*-test analyses.

Studies were performed in compliance with the NCRI Guidelines for the welfare and use of animals in cancer research (Workman *et al.*, 2010) and with licenses issued under the UK Animals (Scientific Procedures) Act 1986 (PPL 40/3212) following local Ethical Committee review.

2.1 Tumor implantation and monitoring

LoVo and HCT116 colorectal carcinoma cells were cultured in RPMI 1640 medium supplemented with 10% heat inactivated fetal calf serum (FCS) at 37°C in a humidified 5% CO_2 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 5×10^6 cells per mouse (in 0.1 mL serum-free culture medium) injected subcutaneous in female nu/nu CBA mice aged 10 weeks old.

Tumor size was monitored using callipers and the formula for ellipsoid volume, $V = (\pi/6)LWD$, where L, W and D are the largest orthogonal dimensions of the ellipsoid. When tumors reached 300–400 mm^3 in size, mice were randomized to sham or given tumor-localized RT (single 10 Gy fraction) using a metal-ceramic MXR-320/36 X-ray machine (320 kV, Comet AG, Switzerland). The RT was administered under ambient conditions to restrained, non-anesthetized 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 h post RT along with calliper measurement of tumor volume. After the second MRI scan, animals were killed humanely by cervical dislocation without recovery from anesthesia.

2.2 MRI acquisition and analysis

Mice were anesthetized with isoflurane delivered through a nose cone apparatus at 2 ml/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 7 T 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 min per animal.

Diffusion-weighted imaging (TR/TE = 2250/20 ms; $\alpha = 90°$; b values 150, 500 and 1000 s/mm^2 along one diffusion direction; matrix 128 × 128 and FOV 2.56 × 2.56 cm; 15 contiguous slices of 0.6 mm thickness) was performed after localization with a T2-weighted anatomical sequence (TR/TE = 2410/50; $\alpha = 136.8°$; matrix 256 × 256 and FOV 2.56 × 2.56 cm; 15 contiguous slices of 0.6 mm thickness). ADC maps were generated by selecting a region of interest on the lowest b value image. Voxel-wise values of ADC (Supplementary data file) were calculated using in-house software across the tumor using a least squares fitting routine for the equation $S = S0e^{-bD}$, where $S0$ represents the signal intensity in the absence of a diffusion sensitizing gradient, S is the signal intensity for a particular b value, b is the numerical value in s/mm^2 and D is the apparent diffusion coefficient (mm^2/s).

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°C (Doblas *et al.*, 2015).

Baseline and change in tumor volume and ADC (mean value and IQR) parameters were compared between control and treated tumors using Student's *t*-test for independent samples in IBM SPSS Statistics v.22 (Armonk, NY). All tests were two-tailed. These tests were performed and combined to provide comparison with the statistics derived from LPM (see below). In all tests, $P < 0.05$ was considered to indicate statistical significance. Corrections for multiple comparisons were applied where necessary.

2.3 LPM of ADC data

A linear Poisson model describes a set of histograms (i.e. ADC distributions) using a linear combination of PMFs, where each PMF represents some sub-component (e.g. a mode of variability/behavior) of the signal:

$$\mathbf{H}(\mathrm{ADC}, t) \approx \sum_{C}^{N_C} P(ADC, t|C)\mathbf{Q}_C + \sum_{T}^{N_T} P(ADC, t|T)\mathbf{Q}_T, \quad (1)$$

where $\mathbf{H}(\mathrm{ADC}, t)$ is the histogram bin recording the frequency of observed ADC values within range ADC at time t; C is a label indicating a component of control behavior; T is a label indicating a component of treatment behavior, as determined by the *additional variability* within the treatment group, i.e. the behavior in treated cases that cannot be accounted for already by control behavior; $P(ADC, t|C)$ and $P(ADC, t|T)$ are the probabilities of observing an ADC value in range at time t from control behavior or treatment behavior; and $\mathbf{Q}_C$ and $\mathbf{Q}_T$ are the quantities of each component in the data. There are N_C control components and N_T treatment components. Each component can broadly be considered as a type of tissue development in control or treatment, corresponding to a mode of heterogeneous variability. The more complex a tumor and its

response to treatment, the greater the number of components the tumor needs in its model.

Given a set of control tumors, $i \in \{1, 2, \ldots, S_C\}$ (where S_C is the control cohort sample size) and treated tumors, $j \in \{1, 2, \ldots, S_T\}$ (where S_T is the treatment cohort sample size), an LPM is used to provide likelihood solutions to PMFs and quantities. Estimation of quantities and probabilities are achieved using expectation maximization to optimize the following extended maximum likelihood for control cohorts:

$$\ln \mathcal{L} = \sum_{i,ADC,t} \ln\left[\sum_{C}^{N_C} P(ADC,t|C)\mathbf{Q}_{Ci}\right]\mathbf{H}_i(ADC,t) - \sum_{C}\mathbf{Q}_{Ci} \quad (2)$$

and the following for treatment cohorts:

$$\ln \mathcal{L} = \sum_{j,ADC,t} \ln\left[\sum_{C}^{N_C} P(ADC,t|C)\mathbf{Q}_{Cj} + \sum_{T}^{N_T} P(ADC,t|T)\mathbf{Q}_{Tj}\right] \quad (3)$$

$$\times \mathbf{H}_j(ADC,t) - \sum_{C}\mathbf{Q}_{Cj} - \sum_{T}\mathbf{Q}_{Tj}.$$

Thus, the model is trained in two parts. Initially, N_C terms are estimated using only the control cohort as training data [Equation (2)]. Once the PMFs for control behavior have been learnt, these components are automatically included as modes of behavior within the treatment cohort. The additional N_T components that describe the extra variability expected due to treatment are then learnt using the treatment cohort, keeping the original N_C components as part of the model [Equation (3)]. In this way, parts of a treated tumor's ADC distribution can be partitioned into quantities of responding and non-responding behavior.

A model selection process identifies the optimum number of LPM components required to describe the ADC distributions. Multiple models are constructed with increasing numbers of components, with the best fitting models being selected for use in subsequent analysis. The number of components required to describe each class of response, i.e. N_C for control and $N_C + N_T$ for treatment, is determined by adding additional components until the χ^2 per degree of freedom between LPM and ADC histograms reaches a minimum, ideally at unity:

$$\chi^2_D = \frac{1}{D}\sum_{\mathrm{ADC},t}\frac{\left[\sqrt{\mathbf{H}(\mathrm{ADC},t)} - \sqrt{\mathbf{M}(\mathrm{ADC},t)}\right]^2}{\sigma^2_{\mathrm{ADC},t}}, \quad (4)$$

where D is the number of degrees of freedom and σ^2 is the variance predicted on the residual. The square-roots are present to transform the Poisson distributed histogram frequencies into Gaussian-like variables to improve this figure of merit's approximation to ideal χ^2 statistics, as described in Anscombe (1948).

Assuming independent Poisson errors ($\sigma^2_{\mathbf{H}} \approx \mathbf{H}$), LPMs provide estimates of uncertainties by summing the effects of individual Poisson bins into quantity error covariances. This is achieved using error propagation. The error covariance can be further scaled by χ^2_D (goodness-of-fit) computed from LPM-data residuals to boost errors to better match actual distributions of true residuals, i.e. scaling factor that can be caused by the up-sampling of MRI data. A covariance matrix for quantities, $\mathbf{Q}$, can be estimated using

$$\mathbf{C}_{ij} = \chi^2_D \sum_{m}\left[\left(\frac{\partial \mathbf{Q}_i}{\partial \mathbf{H}(ADC,t)}\right)\left(\frac{\partial \mathbf{Q}_j}{\partial \mathbf{H}(ADC,t)}\right)\sigma^2_{\mathbf{H}}\right], \quad (5)$$

where $\mathbf{C}$ is the error covariance matrix for the estimated quantities.

Table 1. Conventional *t*-test analysis

Measurement	*Z* score	*P*-value
LoVo vol. change	3.3	0.001
LoVo mean ADC change	3.5	0.0004
LoVo IQR change	2.0	0.041
LoVo combined	**5.2**	**<0.000001**
HCT116 vol. change	4.6	0.0008
HCT116 mean ADC change	4.3	0.0009
HCT116 IQR change	2.1	0.047
HCT116 combined	**8.1**	**<0.000001**

Note: The cohort level significances are 5.2 SD change and 8.1 SD change for LoVo and HCT116, respectively. These figures should be compared to the cohort level significances of Tables 2 and 3.

The statistical significance of treatment response is computed by dividing the sum of treatment quantities $\sum_T \mathbf{Q}_T$ by the estimated error on that total quantity. This provides a Z-score, indicating how many standard deviations from zero the response is estimated to be.

2.4 Model validation

The null hypothesis from which *P*-values are computed is that behavior is consistent with control, and that control behavior is predictable. This behavior must generalize to unseen control data. In contrast, treatment behavior only needs to be different from control. The validity of this null hypothesis relies upon there being no significant changes in independent non-treatment groups. We used a combination of control and leave-one-out testing to provide technical validation.

Treatment models were fitted to control training data to ensure the measured effects of treatment were consistent with zero (with error bars). Additionally, if control LPM is representative of typical non-treated tumors, 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 therefore 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 datasets that are typical of those used in preclinical cancer imaging experiments (Bernsen *et al.*, 2015). This approach also protects against false-positive results through quality control (i.e. representativeness testing) of training data.

3 Results

3.1 Cohort volumetrics and summary ADC detect RT response

Volume and basic ADC distribution parameters demonstrated that significant growth inhibition was induced by RT in both xenograft models at 72 h, relative to control. In both LoVo and HCT116, RT reduced volume, increased mean ADC value and increased IQR of the ADC distribution, relative to control. Treatment effects were detected at the *cohort level*, as summarized in Table 1, reaching high levels of significance ($P < 0.000001$). The Z-scores and *P*-values were computed from *t*-tests on the three parameters individually (changes in volume, mean and IQR). The combined Z-score and *P*-value values show the significances attainable when the three parameters are considered jointly, assuming each provides independence evidence of change. As *t*-tests are applied to the group, individual tumor change assessments are not possible using this method.

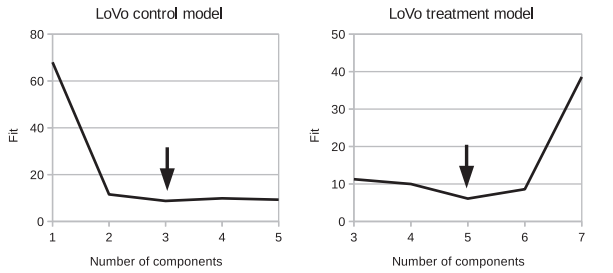

Fig. 2. Model selection curves indicating necessary number of components to describe control and treatment groups. Left: χ^2_D as a function of N_C for LoVo. Right: χ^2_D as a function of $N_C + N_T$ for LoVo

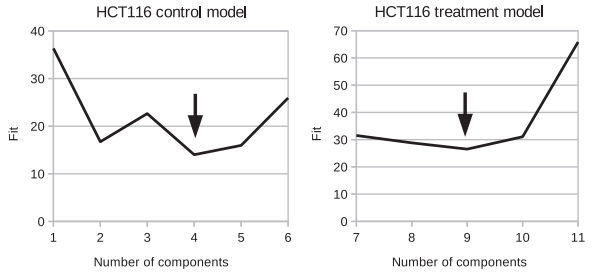

Fig. 3. Model selection curves indicating necessary number of components to describe control and treatment groups. Left: χ^2_D as a function of N_C for HCT116. Right: χ^2_D as a function of $N_C + N_T$ for HCT116

3.2 LPM identifies the varying complexity of different xenograft models

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 three components to describe ADC distributions in the LoVo control tumors, with an additional two required for the variability caused by treatment. An equivalent and independent process was performed for the HCT116 tumors. This yielded four components in control tumors and an additional five for treatment response.

The plots in Figures 2 and 3 show these results in detail in terms of goodness-of-fits (χ^2_D) for models with different numbers of components. We seek the number of components that gives the minimum. The best solutions are indicated with arrows: LoVo $N_C = 3$ and $N_T = 2$ ($N_C + N_T = 5$ in the plot); HCT116 $NC = 4$ and $N_T = 5$ ($N_C + N_T = 9$ in the plot).

The HCT116 tumors are expected to be more complex than LoVo, as they show a greater inter-quartile range of ADC values and can be seen to be more heterogeneous upon visual inspection. A more complex tumor is expected to require a greater number of LPM components to be modeled. The LPM data indicate that the HCT116 xenografts were more spatially complex than the LoVo xenografts and that LPM can detect this differing level of tumor complexity, which is expected of these particular tumors. The greater number of components required to describe HCT116 tumors reflects the higher variability that can be seen visually in Figure 1.

The ADC distributions associated with the extracted components can be seen in Figures 4 and 5. Each component is a probability distribution, showing the statistical correlations between ADC values between the two time points. These correspond to the $P(ADC, t|T)$ and $P(ADC, t|C)$ parts of the model. The weighted sum of these distributions describes the variability observed within the data. Biologically, each component can be interpreted as a sub-population

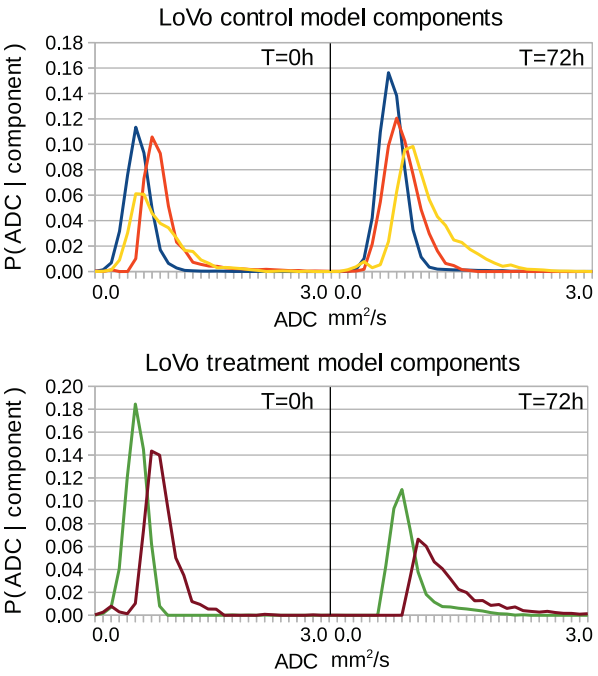

Fig. 4. Estimated components (PMFS: $P(ADC, t|C)$ and $P(ADC, t|T)$), one color per component. Left and right plots indicate baseline and 72 h. Top: LoVo control components. Bottom: LoVo treatment components

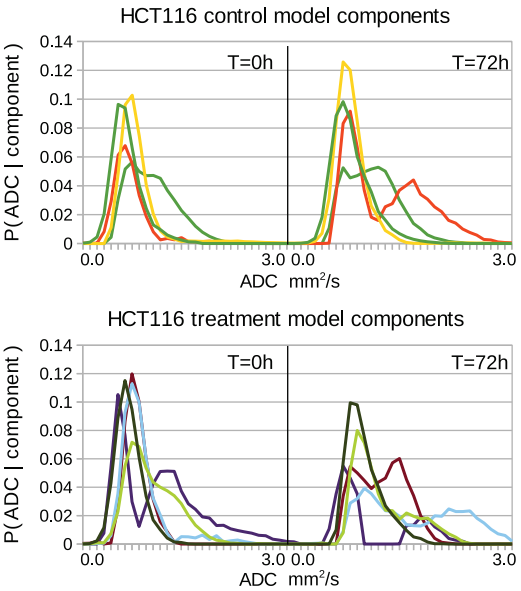

Fig. 5. Estimated components (PMFS: $P(ADC, t|C)$ and $P(ADC, t|T)$), one color per component. Left and right plots indicate baseline and 72 h. Top: HCT116 control components. Bottom: HCT116 treatment components

of ADC values found within the tumors. The higher ADC values at $t = 72$ are more probable, indicating greater diffusion due to less restricted fluid movement.

3.3 LPM validation identifies outliers in control groups

We used control testing and a leave-one-out approach to validate the ability of the model to distinguish data with different ADC distributions to ensure control growth was corrected accounted for.

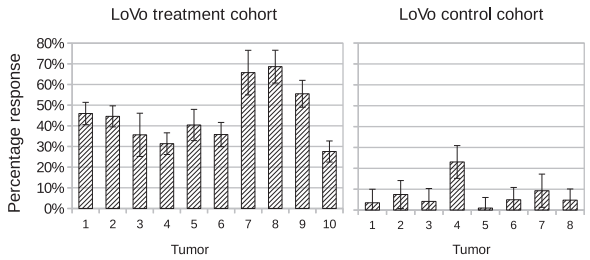

Fig. 6. Volume response to treatment (i.e. $\sum_T Q_T$) for LoVo tumors. Left: Treatment cohort, with significant non-zero values. Right: Control cohort, with values consistent with zero (i.e. within level of predicted error) with possible outlier at tumor 4. All error bars show ± 1 SD

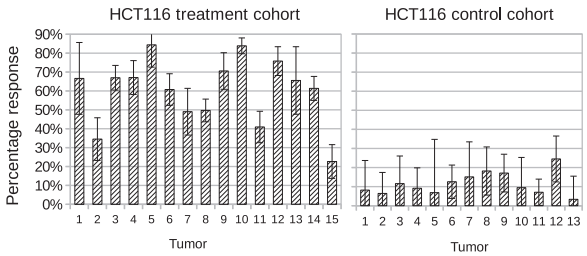

Fig. 7. Volume response to treatment (i.e. $\sum_T Q_T$) for HCT116 tumors. Left: Treatment cohort, with significant non-zero values. Right: control cohort, with values consistent with zero (i.e. within level of predicted error). All error bars show ± 1 SD

To do this, we applied fully trained models (i.e. leave-all-in) to both LoVo and HCT116 control data, followed by reduced models where each tumor in turn is excluded before being used as an independent test data point (i.e. leave-one-out). Responses were computed in each case with leave-all-in results plotted in the right of Figures 6 and 7. Leave-all-in and leave-one-out results are directly compared in Tables 4 and 5.

A Z score is given by dividing the size of a response by 1 SD error on that response. In cohorts of around 10, all control tumors would be expected to have Z scores of <2. This was found for all but two tumors (LoVo 4 and HCT116 12), with average Z scores from full models of 1.05 for LoVo and 0.94 for HCT116. 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. 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. During full analysis LoVo control tumor 4 showed a Z score of 2.9 for estimated treatment volume and HCT116 control tumor 12 showed a Z score of 2.0. These increased to 5.2 and 3.5, respectively, for leave-one-out analysis, implying differences from other control data. This could be explained by the data being an atypical, yet otherwise valid, control sample, which could have been better modeled 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.

3.4 LPM quantifies the percentage responding volume in each tumor

Non-responding tumor was defined by the sum of the control model component volumes ($\sum_C Q_C$) and responding tumor was defined by the sum of the treated model component volumes ($\sum_T Q_T$). The proportion of tumor changing with therapy was calculated, along with error bars (right of Figs 6 and 7). All LoVo and HCT116 tumors treated with RT showed statistically significant volumes of responding tumor, i.e. the responding volume was above zero, beyond the level expected by noise alone. For LoVo, proportion of volume responding to RT varied between 27.6 and 68.6% (median responding volume 40.4%). For HCT116, proportion of volume responding to RT varied between 22.7 and 84.4% (median responding volume 61.4%). In comparison, all control tumors (except outlier LoVo control tumor 4) had responding volumes consistent with zero. These measurements are possible with the LPM method, but not the t-test method.

3.5 LPM biomarkers of response are more powerful than conventional analyses

In LPM, the error estimates on measured affected volumes incorporate systematic processes associated with learning the model parameters (i.e. determination of PMFs), as well as the statistical errors on weighting factors used to describe each case. 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 datasets, by testing the null hypothesis (i.e. zero response) on a case by case basis. The probability of the treatment volume being consistent with zero on the basis of estimated error was measured. Tables 2 and 3 show the individual and cohort significances and Tables 4 and 5 show the control cohort responses for comparison.

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 for conventional volume and ADC (mean and IQR), giving a total Z score of 5.2 SD for LoVo (Table 1). 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 SD for LoVo tumors. 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. This equates to an increase in power of approximately 4-fold. An equivalent comparison of summary statistics and LPM statistics in HCT116 xenografts treated with RT showed a similar gain in statistical power. These data reveal that mathematical modeling of imaging data through LPM enables substantial increase in statistical power to detect response to therapy.

4 Discussion

In this study, we describe how modeling the spatial heterogeneity present in imaging data can increase statistical power of identifying response to therapy. We investigated a technique called LPM in a

Table 2. LoVo treatment cohort result significances

Tumor	*Z* score	*P*-value	Effect (%)	Error (%)
1	8.5	<0.000001	45.98	5.40
2	8.8	<0.000001	44.59	5.09
3	3.4	0.000667	35.64	10.47
4	5.9	<0.000001	31.37	5.24
5	5.3	<0.000001	40.41	7.57
6	6.1	<0.000001	35.77	5.87
7	6.1	<0.000001	65.75	10.80
8	8.6	<0.000001	68.64	7.94
9	8.5	<0.000001	55.51	6.50
10	5.4	<0.000001	27.57	5.10
Combined	**21.8**	**<0.000001**		

Note: The cohort-level significance (bottom row) is approximately four times that for LoVo in Table 1.

Table 3. HCT116 treatment cohort result significances

Tumor	*Z* score	*P*-value	Effect (%)	Error (%)
1	3.5	0.000453	66.66	19.01
2	3.1	0.002239	34.53	11.30
3	10.3	<0.000001	67.00	6.53
4	7.5	<0.000001	67.10	8.97
5	7.2	<0.000001	84.41	11.80
6	7.3	<0.000001	60.77	8.35
7	3.9	0.000072	49.04	12.36
8	8.3	<0.000001	49.71	6.01
9	7.3	<0.000001	70.58	9.67
10	20.1	<0.000001	83.89	4.18
11	4.9	<0.000001	40.94	8.32
12	9.9	<0.000001	75.82	7.61
13	3.7	0.000243	65.56	17.87
14	9.7	<0.000001	61.36	6.35
15	2.5	0.0111474	22.67	8.93
Combined	**32.6**	**<0.000001**		

Note: The cohort-level significance (bottom row) is approximately four times that for HCT116 in Table 1.

Table 4. LoVo control cohort result significances

Tumor	*Z*	(*Z*)	*P*	(*P*)	Effect (%)	(%)	Error (%)	(%)
1	0.5	(1.2)	0.58	(0.20)	3.51	(23.92)	6.48	(18.86)
2	1.1	(1.0)	0.26	(0.31)	7.47	(8.89)	6.69	(8.75)
3	0.6	(0.5)	0.49	(0.57)	4.17	(5.95)	6.18	(10.73)
4	2.9	(5.2)	0.00	(0.00)	23.05	(32.73)	7.84	(6.29)
5	0.2	(0.0)	0.84	(0.97)	0.98	(0.38)	5.07	(11.83)
6	0.8	(0.3)	0.40	(0.72)	4.95	(9.24)	5.89	(26.15)
7	1.2	(0.4)	0.24	(0.68)	9.28	(6.10)	7.96	(15.17)
8	0.9	(0.5)	0.37	(0.56)	4.78	(7.99)	5.34%	(13.76)

Note: Main figures show results for leave-all-in analysis. Figures in brackets show leave-one-out results, where the model was trained on all except the current tumor before being applied to the current tumor.

Table 5. HCT116 control cohort result significances

Tumor	Z	(Z)	*P*	(*P*)	Effect (%)	(%)	Error (%)	(%)
1	0.5	(0.1)	0.59	(0.86)	8.18	(10.29)	15.53	(59.55)
2	0.6	(0.3)	0.55	(0.70)	6.43	(7.41)	11.00	(19.52)
3	0.8	(1.0)	0.42	(0.29)	11.57	(23.31)	14.48	(22.07)
4	0.9	(0.7)	0.39	(0.44)	9.14	(7.97)	10.70	(10.53)
5	0.2	(0.3)	0.80	(0.71)	6.84	(13.61)	27.93	(36.93)
6	1.4	(1.8)	0.15	(0.58)	12.55	(29.65)	8.74	(15.65)
7	0.8	(0.5)	0.41	(0.56)	15.10	(16.67)	18.36	(29.09)
8	1.4	(2.2)	0.14	(0.02)	18.18	(36.09)	12.62	(16.05)
9	1.7	(1.4)	0.08	(0.16)	17.09	(20.52)	9.89	(14.62)
10	0.6	(0.8)	0.54	(0.37)	9.45	(13.43)	15.84	(15.19)
11	1.0	(0.8)	0.30	(0.39)	7.04	(20.15)	6.89	(23.91)
12	2.0	(3.5)	0.04	(0.00)	24.50	(36.00)	12.01	(10.10)
13	0.3	(0.0)	0.78	(0.96)	3.34	(2.91)	12.14	(69.60)

Note: Main figures show results for leave-all-in analysis. Figures in brackets show leave-one-out results, where the model was trained on all except the current tumor before being applied to the current tumor.

well understood biological paradigm, namely ADC as a response biomarker following high-dose RT.

Next, we demonstrated that LPM could appropriately describe ADC distributions of varying complexity, across two untreated xenograft models, with multiple model components being determined to account for modes of tumor heterogeneity. 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.

First, 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 (de Jong *et al.*, 2014). In the era of personalized medicine that employs tumor models of increasing biological relevance and complexity (Sharpless and Depinho, 2006), 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.

Second, any pair (pre- and post-) of ADC values can be assigned a probability (p-value or Z-score) 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 4-fold in HCT116.

Third, this feature enables response detection on a sample by sample basis, without the need for spatial mapping, e.g. image segmentation and pre- post- treatment coregistration. This is possible since LPM models variation within control data and then can account for this in the treatment group, identifying the number of voxels that are different within the frequency distribution of data, as opposed to the spatial distribution. 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, equating to a 4-fold increase in power.

The implications of these data are substantial. Once a control 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 would allow reduction in animal numbers, with welfare benefits (Workman *et al.*, 2010), and the 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 (Malaney *et al.*, 2014) and in co-clinical trials where multiple therapies are tested against animal models with different genetic knockdown/knockout features (Clohessy and Pandolfi, 2015).

The automatic process of building an individual LPM model and computing its errors takes <5 min on typical hardware. However, the process must be performed multiple times during model selection and validation. The model selection process for LoVo required 10 models to be constructed, whereas HCT116 required 11. Leave-one-out validation required an additional eight models for LoVo and 13 for HCT116, representing each possible leave-one-out control combination. Total run-time was <4 h, making it feasible to perform multiple complete analyses per day.

The LPM method described here has some limitations. As the volume of responding tissue is computed by excluding all variation that 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 labeling 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 (O'connor, 2017). 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 (Doblas *et al.*, 2015). 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 χ^2_D, and since quantity error covariances are scaled by this value, the statistical significance of treatment estimates is penalized for poorly modeled data. Large quantity errors can generally be attributed to poor models, for example, with few control datasets, 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 penalizing 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.

5 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.

Acknowledgements

We woud like to thank Isabel Peset, Francesca Trapani, Garry Ashton, Caron Abbey and Steve Bagley of the Cancer Research UK Manchester Institute, University of Manchester for additional support.

Funding

This work was supported by the Leverhulme Trust funding [grant number RPG-2014-019] to N.A.T., Royal College of Radiologists Small Project Grant to J.P.B.O., Cancer Research UK (CRUK) Clinician Scientist award [grant number C19221/A22746] to J.P.B.O. and CRUK and EPSRC Cancer Imaging Centre in Cambridge and Manchester funding to The University of Manchester [grant number C8742/A18097] to N.A.T., K.J.W. and J.P.B.O.

Conflict of Interest: none declared.

References

Alizadeh,A. *et al.* (2015) Toward understanding and exploiting tumor heterogeneity. *Nat. Med.*, **21**, 846–853.

Anscombe,F. (1948) The transformation of poisson, binomial and negative-binomial data. *Biometrika*, **35**, 246–254.

Bedard,P.L. *et al.* (2013) Tumour heterogeneity in the clinic. *Nature*, **501**, 355–364.

Bernsen,M. *et al.* (2015) Biomarkers in preclinical cancer imaging. *Eur. J. Nucl. Med. Mol. Imaging*, **42**, 579–596.

Clohessy,J. and Pandolfi,P. (2015) Pandolfi pp. mouse hospital and co-clinical trial project–from bench to bedside. *Nat. Rev. Clin. Oncol.*, **12**, 491–498.

Comon,P. (1994) Independent component analysis—a new concept? *Sig. Processing*, **36**, 287–314.

Conway,J. *et al.* (2014) Developments in preclinical cancer imaging: innovating the discovery of therapeutics. *Nat. Rev. Cancer*, **14**, 314–328.

de Jong,M. *et al.* (2014) Imaging preclinical tumour models: improving translational power. *Nat. Rev. Cancer*, **14**, 481–493.

Deepaisarn,S. *et al.* (2017) Quantifying biological samples using linear poisson independent component analysis for MALDI-TOF mass spectra. *Bioinformatics*, [Epub ahead of print, doi: 10.1093/bioinformatics/btx630, October 28, 2017].

Dempster,A. (1977) Maximum likelihood from incomplete data via the em algorithm. *J. R. Stat. Soc. Ser. B*, **39**, 1–38.

Doblas,S. *et al.* (2015) Apparent diffusion coefficient is highly reproducible on preclinical imaging systems: evidence from a seven-center multivendor study. *J. Magn. Reson. Imaging*, **42**, 1759–1764.

Gerlinger,M. *et al.* (2012) Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.*, **366**, 883–892.

Gibbs,J. (2000) Mechanism-based target identification and drug discovery in cancer research. *Science*, **287**, 1969–1973.

Gurcan,M. *et al*. (2009) Histopathological image analysis: a review. *IEEE Rev. Biomed. Eng*., **2**, 147–171.

Heppner,G. (1984) Tumor heterogeneity. *Cancer Res*., **44**, 2259–2265.

Just,N. (2014) Improving tumour heterogeneity MRI assessment with histograms. *Br. J. Cancer*, **111**, 2205–2213.

Kibby,M. (1986) Spreadsheet statistics. *Bioinformatics*, **2**, 151–157.

Malaney,P. *et al*. (2014) One mouse, one patient paradigm: new avatars of personalized cancer therapy. *Cancer Lett*., **344**, 1–12.

O'Connor,J. (2017) Cancer heterogeneity and imaging. *Semin. Cell Dev. Biol*., **64**, 48–57.

O'Connor,J. *et al*. (2015) Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome. *Clin. Cancer Res*., **21**, 249–257.

Padhani,A. *et al*. (2009) Diffusion-weighted magnetic resonance imaging as a cancer biomarker: consensus and recommendations. *Neoplasia*, **11**, 102–125.

Sharpless,N. and Depinho,R. (2006) The mighty mouse: genetically engineered mouse models in cancer drug development. *Nat. Rev. Drug Discov*., **5**, 741–754.

Tar,P. and Thacker,N. (2014) Linear poisson models: a pattern recognition solution to the histogram composition problem. *Ann. BMVA*, **2014**, 1–22.

Tar,P. *et al*. (2015) Automated quantitative measurements and associated error covariances for planetary image analysis. *Adv. Space Res*., **56**, 92–105.

Workman,P. *et al*. (2006) Minimally invasive pharmacokinetic and pharmacodynamic technologies in hypothesis-testing clinical trials of innovative therapies. *J. Natl. Cancer Inst*., **98**, 580–598.

Workman,P. *et al*. (2010) Guidelines for the welfare and use of animals in cancer research. *Br. J. Cancer*, **102**, 1555–1577.

Xia,J. *et al*. (2017) An ensemble approach to protein fold classification by integration of template-based assignment and support vector machine classifier. *Bioinformatics*, **33**, 863.

Zeng,T. *et al*. (2017) Deepem3d: approaching human-level performance on 3d anisotropic em image segmentation. *Bioinformatics*, **33**: 2555–2562.

Zong,N. *et al*. (2017) Deep mining heterogeneous networks of biomedical linked data to predict novel drug-target associations. *Bioinformatics*, **33**: 2337–2344.