# scientific reports

Check for updates

## OPEN

# Combining multiple spatial statistics enhances the description of immune cell localisation within tumours

Joshua A. Bull[1]✉, Philip S. Macklin[2], Tom Quaiser[3], Franziska Braun[3], Sarah L. Waters[4], Chris W. Pugh[2] & Helen M. Byrne[1]

Digital pathology enables computational analysis algorithms to be applied at scale to histological images. An example is the identification of immune cells within solid tumours. Image analysis algorithms can extract precise cell locations from immunohistochemistry slides, but the resulting spatial coordinates, or point patterns, can be difficult to interpret. Since localisation of immune cells within tumours may reflect their functional status and correlates with patient prognosis, novel descriptors of their spatial distributions are of biological and clinical interest. A range of spatial statistics have been used to analyse such point patterns but, individually, these approaches only partially describe complex immune cell distributions. In this study, we apply three spatial statistics to locations of CD68+ macrophages within human head and neck tumours, and show that images grouped semi-quantitatively by a pathologist share similar statistics. We generate a synthetic dataset which emulates human samples and use it to demonstrate that combining multiple spatial statistics with a maximum likelihood approach better predicts human classifications than any single statistic. We can also estimate the error associated with our classifications. Importantly, this methodology is adaptable and can be extended to other histological investigations or applied to point patterns outside of histology.

The importance of the immune response to tumour biology and therapy is increasingly recognised and improved understanding is required. Immune responses are multifaceted, with different functions being mediated by individual leukocyte sub-types. These leukocyte sub-types (and some aspects of their function) are defined by the expression of unique proteins. The location within tumour material of these proteins, and thus the cells expressing them, can be identified by immunohistochemistry[1]. Since approximately 50% of a tumour is composed of host-derived stroma, knowledge of which immune cells infiltrate the regions occupied by malignant cells is key to understanding their function in influencing tumour biology[2-4]. In this study we use as an exemplar the distribution of cells expressing CD68, a protein expressed by macrophages.

In current clinical practice assessment of leukocyte infiltration into tumours is undertaken by highly trained histopathologists using semi-quantitative approaches. Detailed analysis is limited by the effort required to make the measurements, concerns over inter-observer variability and uncertainty about how to handle the manifest biological variability present in tumour material. Despite these issues, current estimates of leukocyte infiltration provide clinically useful information that is both helpful in predicting prognosis and guiding treatment decisions[5-7]. It is likely that deeper analysis would reveal currently untapped information, providing an impetus to develop methods that are more quantitative and can be automated.

Digital pathology provides new opportunities to improve and automate tasks[8,9] such as cell counting[10,11], identifying structures within tumours[12,13], and classification and scoring of routine clinical parameters[14-16]. However, it introduces new challenges related to the size of imaging datasets and variability of staining from slide to slide[17]. While existing open source[18-21] and commercial software alternatives can be used to extract $(x, y)$-coordinates of cell centres, we use a pipeline which allows relatively robust and facile identification of cell locations (described

[1]Wolfson Centre for Mathematical Biology, Mathematical Institute, University of Oxford, Oxford OX2 6GG, UK. [2]Nuffield Department of Medicine, NDM Research Building, University of Oxford, Oxford OX3 7FZ, UK. [3]Roche Pharma Research and Early Development, pRED Informatics, Roche Innovation Center Munich, Munich, Germany. [4]Oxford Centre for Industrial and Applied Mathematics, Mathematical Institute, University of Oxford, Oxford OX2 6GG, UK. ✉email: joshua.bull@maths.ox.ac.uk

1

in the Supplementary Information, Sects. A, B). The availability of $(x, y)$-coordinates for each cell of a given type allows their distribution to be described using spatial statistics. Individual spatial statistics provide quantitative mathematical descriptions which have varying degrees of correlation with histopathological classifications[22–30]. Importantly, we show that by combining spatial statistical descriptions of this data we can automatically derive information which is comparable to that currently reported by a pathologist and more accurate than that obtained using a single spatial statistic.

Differences in tissue architecture, cellular morphology and staining intensity permit manual discrimination between tumour cell nests and adjacent stroma in histology images. However, automated tissue compartmentalisation is not straightforward. Whilst several hand-crafted machine learning approaches have been developed[31–33], this task may be best performed through deep learning[34–36], a technique that is not yet routinely available. Therefore, our statistical method considers only the $(x, y)$-coordinates of the point patterns formed by cell centres and does not depend upon the identification of the tumour/stroma boundary. Furthermore, by using only $(x, y)$-coordinates of cell centres there is also no requirement for multiplex immunohistochemistry or depletion of tumour tissue by labelling adjacent histological sections for other markers, although the methodology described here can be easily adapted to incorporate additional information generated by such approaches.

The approaches we describe provide a route to explore extra information about leukocyte distribution that is not specifically captured by current pathology classifications. Examples could include heterogeneity of infiltration (explored in this report), cellular co-localisation and position relative to tumour micro-environmental measures. By providing precise quantitation, this methodology will allow the underlying biological significance and mechanisms to be explored and hopefully be of both scientific and clinical utility.

## Methodology

**Datasets.** We consider two types of point patterns; cell centroids extracted from human head and neck cancer IHC slides, and computer generated synthetic infiltration patterns. Other tumour types are discussed in the Supplementary Information (Sect. F).

*Head and neck cancer: regions of interest.* A cohort of 16 resected human head and neck tumours was stained to show macrophage locations (CD68+). 4 µm sections were cut from formalin-fixed paraffin embedded tissue blocks of 16 cases of human head and neck squamous cell carcinoma. The sections underwent immunohistochemistry staining on a Leica BOND-MAX automated staining machine (Leica Biosystems). Briefly, sections were deparaffinized, underwent epitope retrieval and endogenous peroxidase activity was blocked with 3% hydrogen peroxide (5 min). Subsequently, sections were incubated with the primary antibody (30 min) followed by post-primary and polymer reagents (8 min each). Next, 3,3′-Diaminobenzidine (DAB) chromogen was applied (10 min) and slides were counterstained with haematoxylin (all reagents included within the BOND Polymer Refine Detection kit, Leica Biosystems, catalogue no. DS9800). The following primary antibody was used during staining: CD68 (mouse monoclonal, clone PG-M1), Agilent Technologies, UK (catalogue no. M087601-2), 1:200 concentration. The positive control sample comprised a section of human tonsil tissue. Stained slides were scanned at ×200 magnification using the NanoZoomer S210 digital slide scanner (Hamamatsu). Whole slide images were reviewed by a pathologist (PSM) who annotated tumour regions and any artefactual changes. Non-overlapping 1.5 mm × 1.5 mm regions of interest (ROIs) were then randomly sampled from within the tumour region on each slide until the region was saturated. This resulted in 549 1.5 mm × 1.5 mm ROIs from across the cohort. Centroids of CD68+ cells were extracted from each ROI using a custom image analysis pipeline (see Supplementary Information, Sects. A, B).

As previously reported[37], patients were approached and informed consent obtained for use of their tissue. Access to these tissue samples for this study was approved under Oxford Radcliffe Biobank (ORB) research tissue bank ethics, reference 09/H0606/5+5 (approved by the National Research Ethics Service [NRES] Committee South Central—Oxford C). All experimental protocols were approved prospectively by the ORB committee and subsequently conducted in accordance with its conditions and those of NRES.

*Synthetic datasets.* The process for generating synthetic data is described in detail in the Supplementary Information, Sect. D. Briefly, we use Gaussian random fields to divide each 1.5 mm × 1.5 mm square into two compartments representing stroma and tumour cell nests. The relative widths of tumour cell nests and stromal structures within the artificial tumour geometries are defined by a characteristic tumour cell nest length scale $l$. An artificial geometry is accepted only if at least 25% of its area is covered by each compartment (tumour or stroma). The number of points in each compartment of the synthetic ROI is controlled to maintain a specified overall cell density $d$ cells per mm$^2$ (the impact of varying $l$ and $d$ on point patterns is explored in Sect. G of the Supplementary Information). The relative density of simulated immune cells in the tumour cell nest ($d_t$) and the stroma ($d_s$), is controlled by a parameter $\rho = \frac{d_t}{d_s}$ which further restricts point placement: low values of $\rho$ generate patterns for which the majority of points are excluded from the tumour cell nest, while values close to 1 generate patterns for which points are uniformly distributed across both regions. Candidate points are sampled randomly from the ROI, and accepted if they do not invalidate these criteria. Points are also rejected if they fall within an exclusion radius of 20 µM of any other point, representing approximately one cell diameter.

The macrophage density across the cohort of head and neck ROIs is approximately normally distributed with mean 333 cells per mm$^2$ and standard deviation 170 cells per mm$^2$. We therefore sample values of $d$ from this distribution when generating synthetic data for testing and validation. We reject distributions having $d < 150$ cells per mm$^2$ to ensure reasonable cell densities when calculating spatial statistics (lower densities are discussed in the Supplementary Information, Sect. E). We construct a training dataset by generating at least 1200 point clouds for values of $\rho$ in the interval $[0, 0.5]$ (increments = 0.02), and $l$ sampled uniformly from the interval
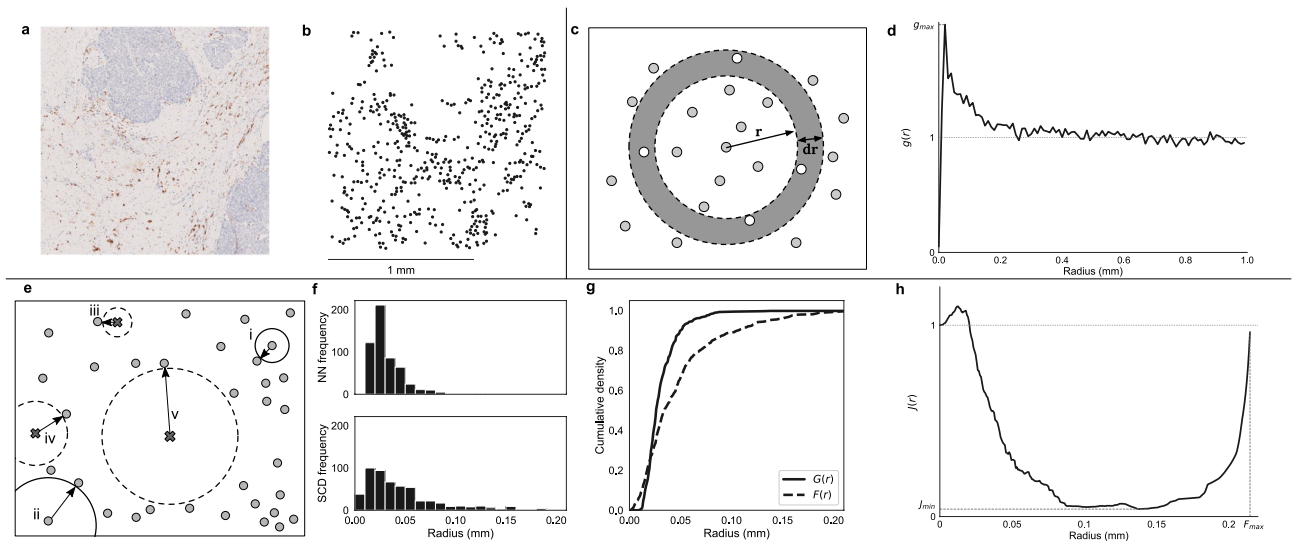
**Figure 1.** (**a**) 1.5 mm × 1.5 mm ROI from the head and neck dataset in which the macrophages (CD68+) are stained brown and the tumour nests pale blue. (**b**) Point cloud of macrophage locations extracted from the image in (**a**). (**c**) Schematic showing calculation of the PCF, $g(r)$. An annulus of width $dr$ and radius $r$ is placed around each point, and the density of points which fall within the annulus is calculated. This density is normalised by the expected density of points within the annulus under complete spatial randomness (CSR) to give $g(r)$. (**d**) The PCF, $g(r)$, for the point cloud in (**b**). (**e**) Schematic illustrating distances used to calculate the nearest neighbour (NN) distribution and spherical contact distribution (SCD); circles = cell centres; crosses = randomly chosen locations. The NN distribution describes the smallest distance between points in the point pattern (circle to circle: i, ii, radius shown as solid line) while the SCD describes distances from randomly chosen locations to points (cross to circle, iii, iv, v, radius shown as dashed line). (**f**) Nearest neighbour distribution (NN) and SCD for the point pattern in (**b**). (**g**) Cumulative distribution functions (CDFs) associated with the histograms in (**f**). The function $G(r)$ is the NN distribution CDF, and $F(r)$ is the CDF of the SCD. (**h**) The $J$-function for the point cloud in (**b**), calculated from the CDFs in (**g**). $F_{max}$ approximates the size of the largest void in the point pattern; it is the largest radius observed in the SCD [see location v in panel (**e**), where the dashed radius is $F_{max}$ for these points]. The $J$-function describes differences in the distribution of points from complete spatial randomness. For a given $r$, smaller values of $J(r)$ suggest clustering, in that more cells will be found in a circle of radius $r$ centred at a cell than in a randomly placed circle. $J_{min}$ describes how densely the cells are clustered across all possible length scales.

[0.1 mm; 0.75 mm]. A separate validation dataset, comprising a further 3680 point clouds, was generated using the same method but with $\rho$ sampled uniformly and at random from the interval [0, 0.5].

**Spatial statistics.** We consider three spatial statistics: the pair correlation function (PCF), the spherical contact distribution (SCD) and the $J$-function. Figure 1 shows typical results when these statistics are applied to a ROI from the head and neck cancer dataset.

*Pair correlation function, g(r).* The PCF (or radial distribution function)[38–40] compares the average density of points against complete spatial randomness (CSR) across length scales. It has previously been used to estimate the length scales of emergent patterns in point clouds representing the locations of distinct cell types[41]. To generate the PCF, an annulus of width $dr$ and radius $r$ is placed around each point (Fig. 1c). The number of points within each annulus is calculated and normalised with respect to the expected number of points that would fall inside the annulus under CSR. This calculation is repeated for each point, and the average value is recorded as a measure of whether points are more or less frequently observed at distance $r$ from a point than expected under CSR. This is repeated for a range of $r$ to generate the PCF, $g(r)$. Figure 1d shows the PCF obtained from the point pattern in Fig. 1b. Further details, including handling of boundary conditions and edge effects, can be found in the Supplementary information, Sect. C. The PCF identifies clustering at distance $r$: $g(r) > 1$ indicates that pairs of points separated by distance $r$ occur frequently, while $g(r) < 1$ indicates that pairs of points are less likely to be separated by distance $r$ than under CSR. In the case of CSR generated through a Poisson process, $g(r) \equiv 1$[40]. In the context of CD68+ macrophages, $g(r) < 1$ indicates that macrophages are rarely observed separated by radius $r$ (occurring only at very small distances, $r \approx 0$ in Fig. 1d. We interpret this as a minimum distance that cell centres must be separated by, corresponding approximately to the radius of a macrophage). $g(r) > 1$ indicates that macrophages are likely to be separated by distance $r$. This occurs for $r < 0.2$ mm in Fig. 1d, showing short-range clustering of macrophages.

*Spherical contact distribution.* The SCD (or empty-space function)[42,43] is closely related to the nearest-neighbour distribution. Figure 1e–g show how the two functions are constructed. The nearest-neighbour distribution

[shown in Fig. 1f for the point pattern in Fig. 1b] is obtained by measuring the distance from each point to its nearest neighbour. The SCD is calculated in a similar way, but reference points are selected randomly from the ROI rather than from the point cloud. For each randomly chosen point, the distance to its nearest neighbour in the point cloud is added to the distribution. Each observation of radius $r$ in the SCD corresponds to a circle of radius $r$ containing no points. Examples of these circular voids can be seen in Fig. 1e, surrounding points iii, iv and v. In the context of CD68+ macrophages, observations at radius $r$ in the SCD therefore indicate the absence of macrophages in a circle of radius $r$ somewhere in the region.

*J-function, J(r).* The *J*-function is a non-parametric test for identifying clustering and dispersion in point patterns[42]. The *J*-function compares the cumulative density functions (CDFs) of the SCD, $F(r)$, and the nearest-neighbour distribution function, $G(r)$, [shown in Fig. 1g for the point pattern in Fig. 1b], where

$$F(r) = \frac{\text{Number of randomly chosen reference points within distance } r \text{ of a target point}}{\text{Number of randomly chosen reference points}}, \tag{1}$$

and

$$G(r) = \frac{\text{Number of points within distance } r \text{ of another point}}{\text{Number of points}}, \tag{2}$$

so that $0 \le F(r) \le 1$ and $0 \le G(r) \le 1$. When calculating $F(r)$ we choose the same number of reference points as there are points in the pattern to ensure that the denominators of Eqs. (1) and (2) are equal. The *J*-function is defined as:

$$J(r) = \frac{1 - G(r)}{1 - F(r)}. \tag{3}$$

Under CSR $G(r) \approx F(r)$, and hence $J(r) \approx 1$. If $J(r) > 1$ then more points are found in a circle of radius $r$ placed randomly in the domain than in a circle centred at one of the points, indicating dispersal of points within the point cloud. If $J(r) < 1$ then a disc of radius $r$ centred on a point contains more points than a randomly placed disc, and we conclude that the points are clustered[42]. Figure 1h shows $J(r)$ for the point pattern in Fig. 1b.

In the context of CD68+ macrophages, $J(r) > 1$ indicates dispersal of macrophages; this is observed at small values of $r$ in Fig. 1h. As with the PCF in Fig. 1d, this indicates that macrophage cell centres are rarely separated by distances of less than 0.02 mm (i.e., the approximate radius of a macrophage.) $J(r) < 1$ indicates clustering of macrophages, which we see at length scales up to approximately 0.2 mm, again correlating with the PCF.

**Summary statistics.** We record three features of the spatial statistics: the peak of the PCF, $g_{max}$; the value $r$ at which $F(r) = 1$, $F_{max}$; and, the minimum value of the *J*-function, $J_{min}$. These features have the following interpretations:

- $g_{max} := \max(g)$. $g(r)$ describes the expected density of points at distance $r$ from another point, compared to CSR. Hence $g_{max}$ describes the maximum intensity of point clustering over all radii $r$.
- $F_{max} := \min(r)$ such that $F(r) = 1$. $F_{max}$ approximates the radius of the largest circular void in the point distribution. As macrophages start to infiltrate into these 'immune deserts' in greater numbers, the voids in the point pattern should reduce in size, and $F_{max}$ will also decrease.
- $J_{min} := \min(J)$. As $J(0) = 1$ and $J \ge 0$, we have $0 \le J_{min} \le 1$. Smaller values of $J_{min}$ indicate that a larger number of macrophages are closer to other macrophages than to randomly selected points (i.e., denser cell clustering).

**Maximum likelihood estimation.** We conduct maximum likelihood estimation (MLE) based on observations of $g_{max}$, $F_{max}$ and $J_{min}$ in order to predict $\rho$ for a given synthetic point pattern. While the likelihood can be estimated directly from the empirical distributions, we instead approximate distributions of $g_{max}$, $F_{max}$ and $J_{min}$ using a normal distribution for each $\rho$. The mean and standard deviation (SD) of these distributions is well approximated using exponential functions of $\rho$ (shown in Fig. 3b–d). This approximation ensures that the maximum likelihood predictions vary monotonically with $g_{max}$, $F_{max}$ or $J_{min}$. This approximation is discussed in the Supplementary Information, Sect. H. As we assume that for each $\rho$ the distributions of $g_{max}$, $F_{max}$ and $J_{min}$ are normal, the log likelihood, $\ln(L)$, can be calculated directly as:

$$\ln L = -0.5 \left( \ln\left(2\pi\sigma^2\right) + \frac{(x - \mu)^2}{\sigma^2} \right) \tag{4}$$

when only one spatial statistic is used, where $\mu$ and $\sigma$ are the mean and SD of the distribution of the statistic, and $x$ is the value of the statistic. When observations from $k$ different statistics are combined, Eq. (4) generalises to give:

$$\ln L = -0.5 \left( \ln|\Sigma| + (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) + k \ln(2\pi) \right) \tag{5}$$

where $\mathbf{x}$ is the vector of observations, $\mu$ is the vector of the means of the distributions and $\Sigma$ is the covariance matrix. While $\mu$ and $\sigma$ are functions of $\rho$ (Supplementary Information, Sect. H), $\Sigma$ must be estimated from the empirical data.

For each $\rho$ we generate a $k$-dimensional grid containing $\ln L$ at evenly spaced values of the spatial statistics. These grids are stacked to form a $k + 1$ dimensional lookup table. For an observation of $k$ spatial statistics, the closest grid points are taken and the profile likelihood is identified from the remaining dimension. The MLE for $\rho$, which we denote $\eta$, is then estimated. A 95% confidence interval around $\eta$ can be found by identifying where the profile log likelihood is above the threshold $\eta - 0.5 \ln(\chi^2(0.95, 1))$, where $\chi^2(0.95, 1)$ is the 0.95 quantile of the chi-squared distribution with 1 degree of freedom[44,45].

## Results

### Application of spatial statistics to histological samples.

We analysed 549 1.5 mm × 1.5 mm regions of interest (ROIs), taken from 16 human head and neck tumour slides stained to show macrophage locations (CD68+). 100 ROIs were randomly selected for manual evaluation by a pathologist as containing "very low", "low", "moderate" or "high" CD68+ macrophage infiltration into tumour cell nests, with 11 subsequently excluded from analysis for containing artefacts such as weak staining or damaged tissue. The pathologist graded CD68+ macrophage infiltration into tumour nests, irrespective of the overall immune cell infiltration within the ROI and without controlling for tumour nest size. CD68+ macrophage point clouds extracted from the remaining 89 ROIs were analysed using the PCF, SCD and $J$-function. Representative ROIs and point clouds are presented in Fig. 2a, while Fig. 2b,c show the PCFs and $J$-functions.

The shapes of the PCF and $J$-function are similar for different ROIs, but properties of the curves change with increasing infiltration (Fig. 2b,c). For each ROI, the PCF is initially low ($g(r) < 1$ when $r < 20 \mu m$), in keeping with volume-exclusion of cell centroids meaning that macrophages are unlikely to be observed within approximately one cell diameter of one another. This is typically followed by a peak ($g(r) > 1$), suggesting short-range clustering.

As $r$ increases the PCF decays towards 1 (grey dotted line), showing decreasing colocalisation between macrophages at larger length scales. The maximum value of the PCF ($g_{max}$, indicated by '*' and dashed lines) decreases with increased macrophage infiltration and decays more rapidly to its asymptotic value. These trends show that increased infiltration implies a more homogeneous distribution of macrophages across a ROI.

The $J$-function reveals similar behaviour (see Fig. 2c). After an initial peak associated with exclusion of points at $r < 20 \mu m$, the $J$-function drops below 1 to a minimum value $J_{min}$ (indicated by '+' and horizontal dashed lines in Fig. 2) and then rises again. $J_{min}$ increases with increasing macrophage infiltration. This suggests that CD68+ cells in ROIs with higher infiltration are less densely clustered than in those with low infiltration. The $J$-function is defined only where $F(r) < 1$; we define this radius as $F_{max}$ (labelled with '*' and vertical dashed lines in Fig. 2c). As $r$ approaches $F_{max}$, the $J$-function becomes infinite (see Fig. 2c), suggesting that the distance between a macrophage and its nearest neighbour is never larger than the radius of immune deserts in the IHC ROIs. ROIs with low infiltration contain large areas with no macrophages, and therefore have higher values of $F_{max}$, than those with high infiltration.

Figure 2d shows that these trends persist across the 89 manually scored ROIs. The CD68+ macrophage density does not correlate with the manual scores. Both $g_{max}$ and $F_{max}$ are negatively correlated with manual scores of increased CD68+ cell heterogeneity, whilst $J_{min}$ is positively correlated. There are insufficient scored samples to identify statistically significant differences between all of the categories, but these trends suggest that high $g_{max}$, high $F_{max}$, and low $J_{min}$ are characteristic of ROIs with low macrophage infiltration. Figure 2e places these statistics into the context of those from the full, unscored dataset. Similarly scored ROIs exhibit similar combinations of spatial statistics, indicating that, with more labelled data, these descriptors could inform a classifier that predicts the manual scores. In practice, manually scoring ROIs is time consuming. Further, it may be impractical to obtain sufficient labelled data to train a classifier. We explain below how, in such circumstances, synthetic point patterns may be used as a surrogate for developing such a classifier.

### Comparison of synthetic data with head and neck ROIs.

Figure 3a compares the head and neck ROIs with synthetically generated training data. The distributions of cell density $d$, $g_{max}$, $F_{max}$ and $J_{min}$ are in good agreement with those calculated from the IHC data. The combinations of spatial statistics observed in the head and neck ROIs are a subset of those in the synthetic data, suggesting that the distribution in a given ROI could be approximated by synthetic data with an appropriately chosen $\rho$. We note that the distributions of spatial statistics in the head and neck data are most similar to synthetic data generated with low $\rho$, suggesting that macrophages tend not to be distributed by CSR. Increasing $\rho$ generates point patterns with spatial statistics similar to those from highly infiltrated ROIs (see Fig. 2). Figure 3a therefore suggests that $\rho$ could be used as a label to describe infiltration in synthetic point patterns. Under this assumption, the synthetic training data could be used to identify relationships between spatial statistics and manual scores of IHC ROIs. Combinations of spatial statistics used to infer $\rho$ for synthetic ROIs could also be used to stratify the IHC samples based on macrophage infiltration.

Figure 3b–d shows how varying $\rho$ affects $g_{max}$, $F_{max}$ and $J_{min}$ in the synthetic training dataset. For fixed $\rho$ observations of $g_{max}$, $F_{max}$ and $J_{min}$ are approximately normally distributed, and the mean and standard deviation (SD) of these distributions are well approximated by exponential functions of $\rho$ (dashed lines in Fig. 3b–d; see Supplementary Information, Sect. H). This suggests that the mean and SD of the distributions of $g_{max}$, $F_{max}$ and $J_{min}$ can be estimated for any value of $\rho \leq 0.5$, including values not present in training data.

### Inference of $\rho$ for synthetic regions.

To test whether $\rho$ can be inferred from observations of $g_{max}$, $F_{max}$ and $J_{min}$, we predict $\rho$ for 3680 point patterns in the synthetic testing dataset, where $\rho$ is distributed uniformly at random across the interval [0,0.5]. We use maximum likelihood estimation based on observations of $g_{max}$, $F_{max}$ and $J_{min}$ to infer the value of $\rho$ used to generate each point pattern. We denote the maximum likelihood estimate
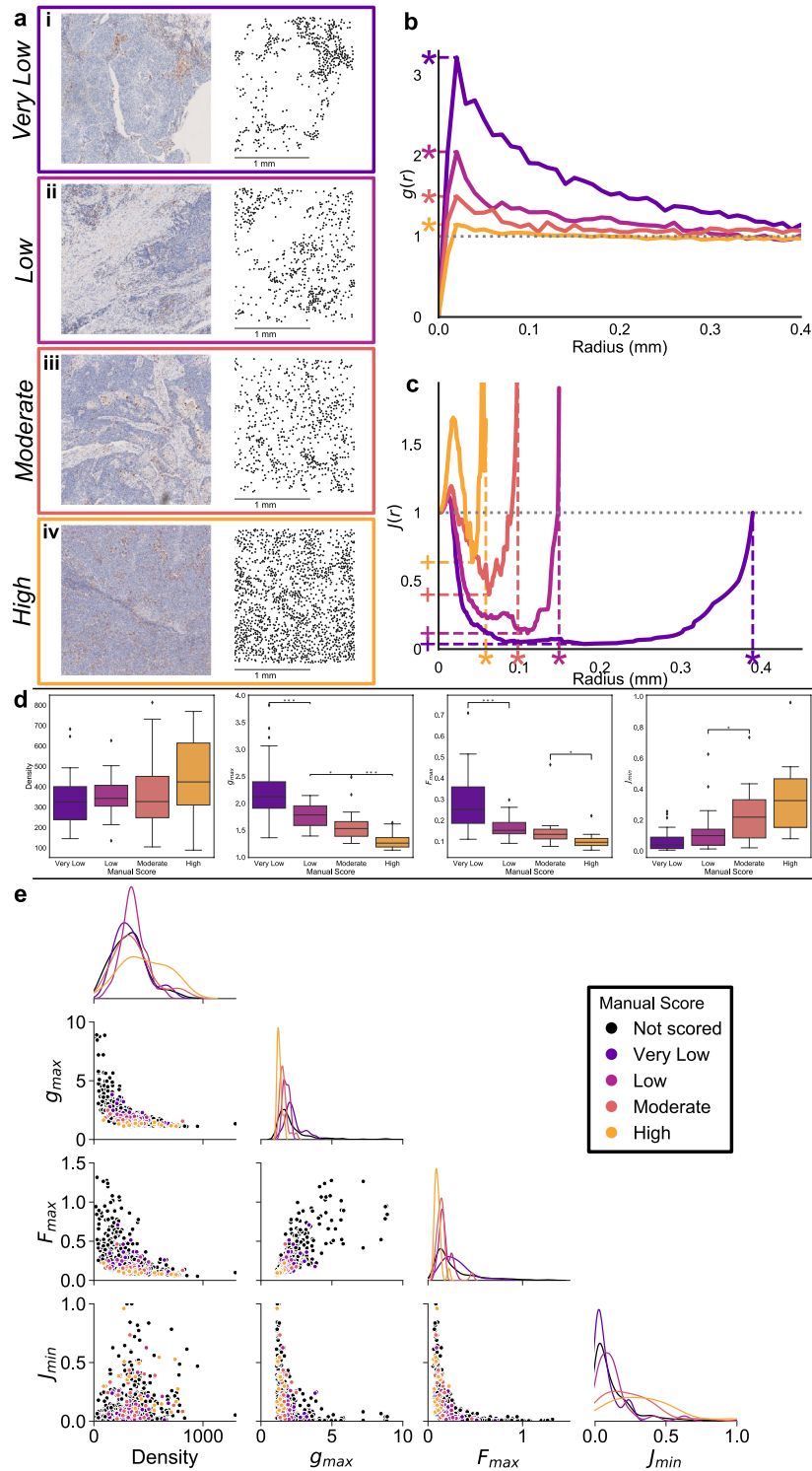
**Figure 2.** (**a**) 1.5 mm × 1.5 mm ROIs from human head and neck tumours assessed as having (i) very low, (ii) low, (iii) moderate, and (iv) high macrophage infiltration into tumour cell nests, with extracted point clouds. (**b**) PCFs for the point clouds in (**a**). As macrophage infiltration increases, the peak of the PCF ($g_{max}$, marked '*') tends towards 1. (**c**) $J$-functions for the point clouds in (**a**). As macrophage infiltration increases, the minimum of the $J$-function ($J_{min}$, marked '+') tends towards 1, while the domain of definition of the $J$-function ($F_{max}$, marked '*') becomes smaller. (**d**) Distribution of $d$, $g_{max}$, $F_{max}$ and $J_{min}$ according to manual scoring of 89 ROIs. $p$ values are the result of $t$-tests on adjacent boxplots. Macrophage density remains consistent across all scoring categories, but $g_{max}$ and $F_{max}$ become smaller with increasing infiltration. $J_{min}$ appears to increase with macrophage infiltration. The differences in $F_{max}$ between low/moderate infiltration, and in $J_{min}$ between very low/low and moderate/high infiltration, are not significant for this dataset. $* : p < 0.05, ** : p < 0.01, *** : p < 0.001$. **e** Distributions (diagonal) and pairwise combinations (off-diagonal) of $d$, $g_{max}$, $F_{max}$ and $J_{min}$ observed in the head and neck dataset. The 89 labelled ROIs are coloured according to manual scoring, and the full unscored dataset of 549 ROIs is shown in black.
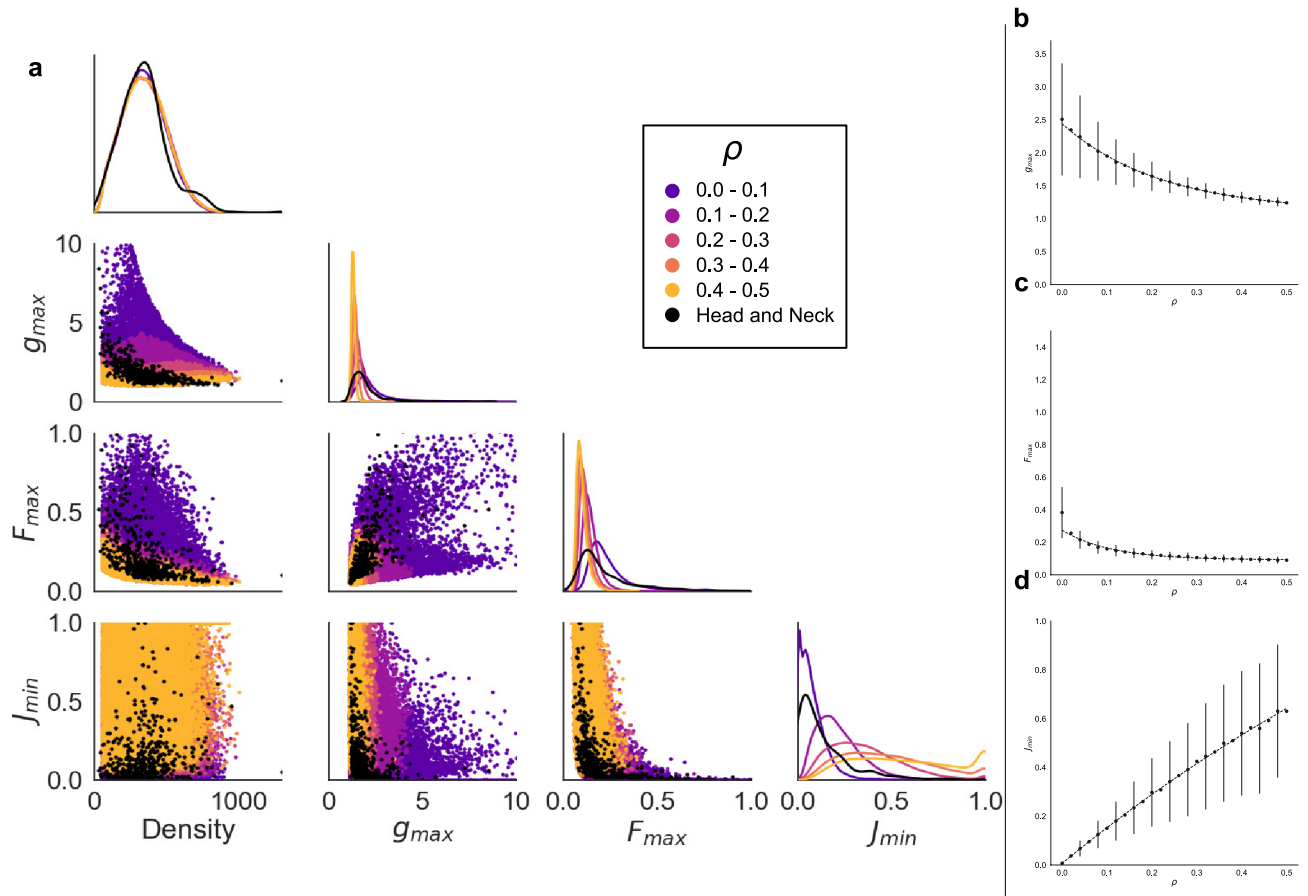
**Figure 3.** (**a**) Distributions (diagonal) and pairwise correlations (off-diagonal) for $d$, $g_{max}$, $F_{max}$ and $J_{min}$. Statistics drawn from the IHC ROIs are shown in black. Other points represent synthetic ROIs in the training dataset and are coloured according to $\rho$. Distributions of cell density are similar for the head and neck data and all ranges of $\rho$ in synthetic data. The statistics extracted from the head and neck data more closely resemble those of synthetic data with low $\rho$ than high. (**b**–**d**) The effect of increasing the degree of infiltration, $\rho$, on (**b**) $g_{max}$, (**c**) $F_{max}$ and (**d**) $J_{min}$ in the synthetic training dataset. Points are the mean of at least 1200 repetitions of randomly generated point clouds. Vertical lines show the standard deviation, and dashed lines show the best fit of the means to an exponential function: (**b**) $g_{max} = 1.34e^{-4.52\rho} + 1.10$, (**c**) $F_{max} = 0.18e^{-9.16\rho} + 0.09$, (**d**) $J_{min} = -2.04e^{-0.74\rho} + 2.06$.

(MLE) $\eta_X$, where $X \subseteq [g, F, J]$ and $g$, $F$ and $J$ indicate the use of $g_{max}$, $F_{max}$ and $J_{min}$ respectively, giving seven possible combinations of spatial statistics to consider. This notation refers to particular instances of MLE prediction, and we will use $\eta$ with no subscript to refer to MLE predicted from any combination of spatial statistics.

Figure 4 shows the MLEs of $\eta$ for different combinations of $g_{max}$, $F_{max}$ and $J_{min}$ for the same point patterns. Each marker corresponds to a different point pattern, and is coloured according to the width of the 95% confidence interval around $\eta$. Predictions in Fig. 4a are based on observations of a single spatial statistic, those in Fig. 4b show predictions of pairwise combinations of the statistics, and Fig. 4c shows predictions made by combining observations of all three statistics. The distance of the predictions to the line $\rho = \eta_X$ can be evaluated using $R^2$. Predictions made based on individual spatial statistics, $\eta_g$, $\eta_F$ and $\eta_J$, are not strongly predictive [$R^2 = 0.5387$, $R^2 = 0.4698$, $R^2 = 0.3033$ respectively, Fig. 4a]. $\eta_g$ generally overestimates $\rho$, but is the most accurate of the single spatial statistics. $\eta_F$ displays banding caused by rounding of observations of spatial statistics (see Supplementary Information, Sect. H). Estimates of $\eta$ close to 0 have small confidence intervals. The width of confidence intervals is also narrow close to $\eta = 0.5$ as the range of possible $\eta$ is restricted to the interval $[0, 0.5]$.

Predictions are improved by combining observations of the spatial statistics (Fig. 4b): the predicted values of $\rho$ are more accurate and the confidence intervals associated with the predictions are smaller and more consistent over a range of values of $\rho$. For example, while $\eta_J$ is the least accurate predictor of the single spatial statistics ($R^2 = 0.3033$), combining observations of $J_{min}$ and $g_{max}$ or $F_{max}$ causes the 95% confidence intervals associated with high values of $\eta$ to become narrower. The most accurate classifier combining two statistics is $\eta_{g,F}$, implying that combining the highest performing spatial statistics may result in a better performing classifier. We note that $\eta_{F,J}$ performs worse than $\eta_F$ and that as such care must be taken when combining observations of spatial statistics to ensure that the resulting classifier is improved. Combining observations of all three statistics in $\eta_{g,F,J}$ yields the highest $R^2$ value ($R^2 = 0.8124$), and confidence intervals which are more consistent across the whole range of $\eta_{g,F,J}$ than for other $\eta$ (mean widths of 95% CI:
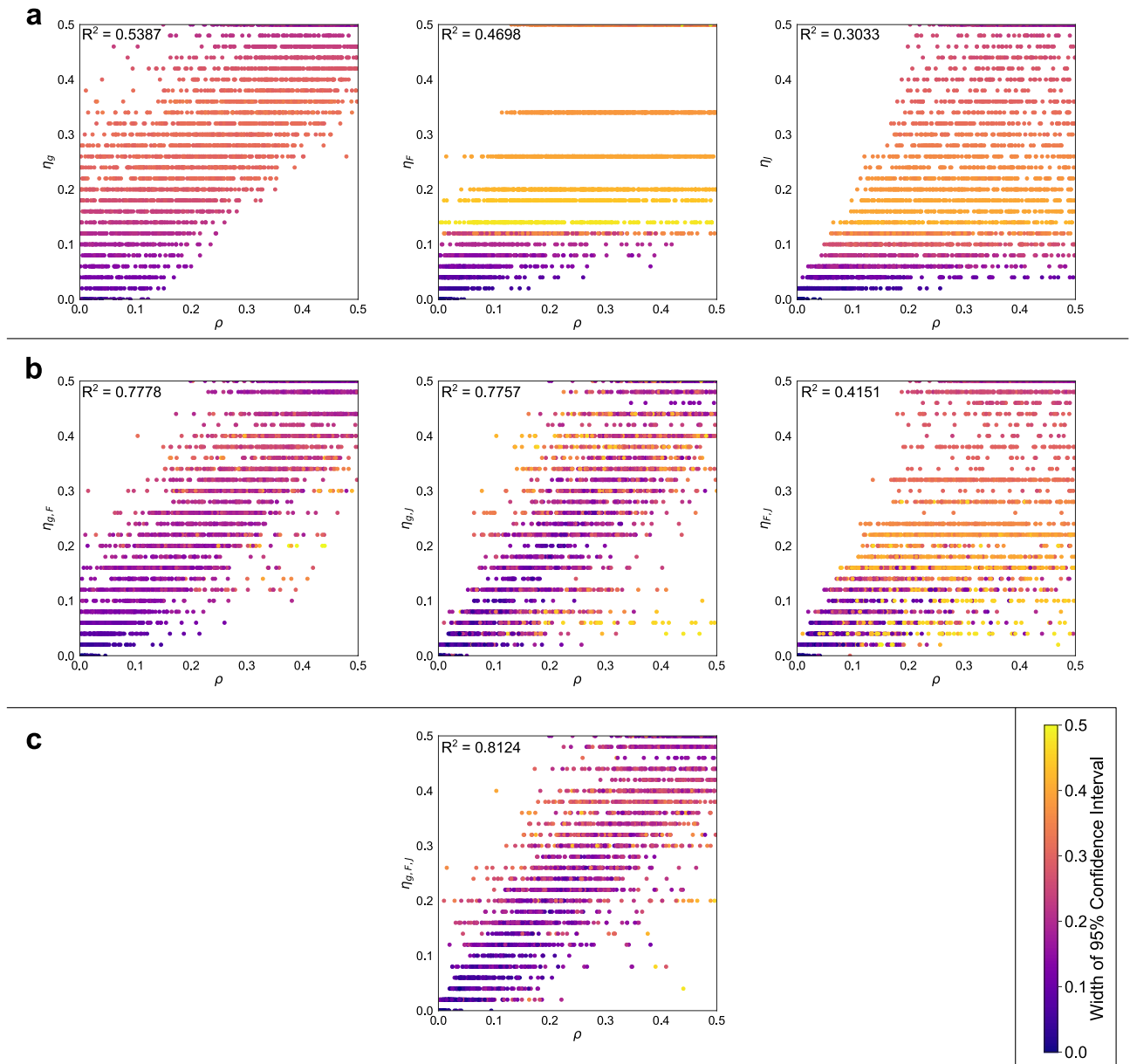
**Figure 4.** Comparison of $\rho$ against $\eta$, for 3680 synthetic point patterns. Points are coloured according to the width of the 95% confidence interval associated with the prediction. Conducting maximum likelihood estimation (MLE) based on observations of multiple statistics can increase prediction accuracy and tighten confidence intervals. $R^2$ values measure deviation from the line $\rho = \eta$. (**a**) $\rho$ based on observations of one spatial statistic. (**b**) $\rho$ based on observations of two spatial statistics. (**c**) $\rho$ based on observations of three spatial statistics.

$\eta_g = 0.256$, $\eta_F = 0.330$, $\eta_J = 0.260$, $\eta_{g,F} = 0.185$, $\eta_{g,J} = 0.197$, $\eta_{F,J} = 0.268$, $\eta_{g,F,J} = 0.171$). Further, observations far from the line $\eta_{g,F,J} = \rho$ typically have wider confidence intervals than more accurately predicted values.

**Automated labelling of histological regions.** The seven metrics $\eta_X$ were applied to the 89 manually assessed head and neck ROIs. Figure 5a–c shows the resulting distributions of $\eta_g$, $\eta_{g,F}$ and $\eta_{g,F,J}$, partitioned by manual score of CD68+ macrophage infiltration. Each classifier identifies statistically significant differences between the manual scoring categories. We note that the pathologist assigns discrete scores to each image whereas $\eta$ is a continuous quantity. Consequently patterns which lie at class boundaries may be difficult to score. This may explain some of the overlap in predictions of $\eta$ between classes. Other extreme predictions are more clearly outliers; the ROIs corresponding to these images are examined in the Supplementary Information, Sect. J.

We note that the ability of $\eta_{g,F,J}$ to distinguish between manual scores is not significantly greater than the ability of $\eta_{g,F}$ or $\eta_g$. However, in line with Fig. 4, the confidence intervals associated with $\eta$ become narrower as observations of additional statistics are combined.
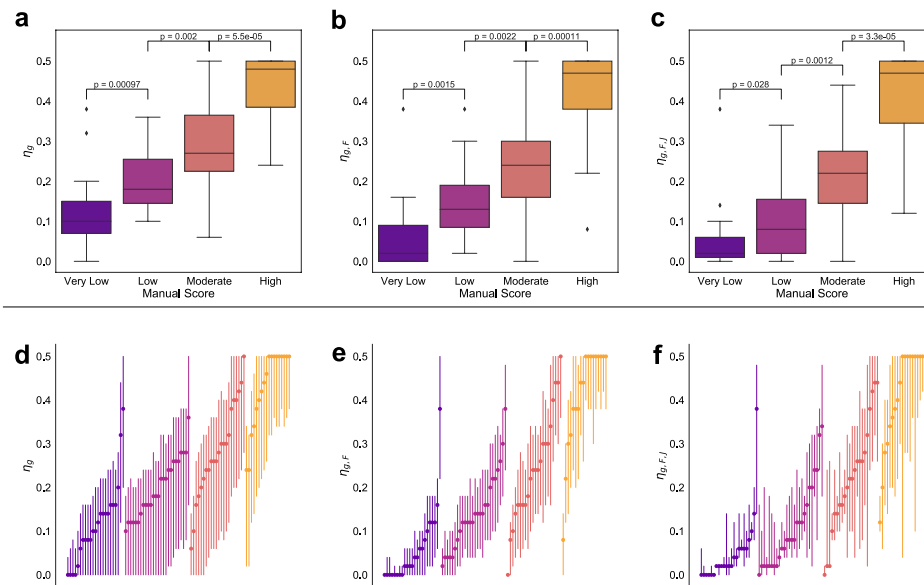
**Figure 5.** (**a**–**c**) Predicted values of $\eta_g$, $\eta_{g,F}$ and $\eta_{g,F,J}$ for the 89 manually scored head and neck ROIs. *p*-values determined using a two-sided *t* test and shown for adjacent boxplots only. (**d**–**f**): 95% confidence intervals (lines) for manually scored head and neck ROIs, for $\eta_g$, $\eta_{g,F}$ and $\eta_{g,F,J}$ (points). In each subplot, ROIs are ordered according to increasing $\eta$, and coloured according to the manual score.

Figure 5d–f shows 95% confidence intervals around $\eta$ for the three classifiers, coloured according to their manual score and ordered in increasing value of $\eta$. While Fig. 5a–c may suggest that $\eta_g$ or $\eta_{g,F}$ most successfully distinguish between different manual categorisations, Fig. 5d indicates that predictions made by $\eta_g$ tend to have wide confidence intervals. Combining statistics can reduce the width of the confidence intervals, as for $\eta_{g,F}$, but as shown in Fig. 4, may also increase their width. Incorporating observations of $J_{min}$ into $\eta_{g,F,J}$ eliminates low values of $\eta$ from the confidence intervals, suggesting that there may be benefits from the inclusion of spatial statistics which do not appear promising in their own right.

A similar analysis for the locations of CD8+ cells within the head and neck cancer dataset can be found in the Supplementary Information, Sect. K.

## Discussion

Although pathologists' training allows them to accurately assess immune cell distributions within tumours, it is infeasible manually to assess large numbers of images in a high throughput manner. Additionally, human assessment is qualitative or, at best, semi-quantitative in nature. Therefore, such a task requires the adoption of carefully calibrated digital image analysis and statistical approaches. Once accurate cell locations have been identified, application of spatial statistics can provide quantitative information about their spatial structure. Whilst individual spatial statistics may identify features such as clustering, it is less clear how these features relate to human descriptions of the complex patterns of immune cell distributions.

In this study, we investigated whether spatial statistics, individually or in combination, agreed with pathological assessment of macrophage distributions within images of human head and neck cancer slides. We identified three summary features ($g_{max}$, $F_{max}$ and $J_{min}$) which vary predictably with the distribution of macrophages. Although each metric correlates with increasing infiltration, there is substantial overlap in the ranges of these metrics and the semi-quantitative categories assigned by the pathologist. We conclude that while observation of a single metric is insufficient to discriminate between semi-quantitative categories assigned by a pathologist, it may be possible to discriminate between them by combining multiple statistics. In statistics and machine learning, including additional parameters in a model may cause overfitting, particularly when the same dataset is used for both testing and validation. While techniques exist to penalise models for overfitting based upon the number of parameters they consider[46,47], in this manuscript the evaluation of the models is performed on an independent test set, which prevents selection bias and means that model complexity need not be explicitly accounted for when determining predictive accuracy. Introducing observations of additional statistics provides a more complete description of the point data, rather than causing overfitting due to increased model complexity.

Since manual assessment of histology images is time-consuming, it is difficult to leverage enough data to enable machine learning approaches. We therefore created synthetic datasets, designed to resemble those encountered in human samples but for which point dispersal is quantified by the parameter $\rho$. We used synthetic training data to determine the probability that a point pattern generated with a specific value of $\rho$ would give rise to particular values of $g_{max}$, $F_{max}$ and $J_{min}$. These probabilities allow estimation of $\rho$ from a given point pattern via MLE, based on observations of one or more of $g_{max}$, $F_{max}$ and $J_{min}$. In this approach, self-weighting ensures that a given metric contributes more to the likelihood function in images where its probability distribution is more

informative. Predictions based on individual spatial statistics were not strongly predictive, but were improved when two spatial statistics were used, and were best when observations of all three statistics were combined. Using multiple statistics also reduced the width of the associated confidence intervals. The ability of MLE to provide an estimate of error for each prediction could also be used to flag images with wide confidence intervals for pathologist review. Furthermore, these descriptors can also flag for human review images with low numbers of cells and/or large regions devoid of immune cells, a common finding in images affected by missing tissue or other histological artefacts (see Supplementary Information, Sect. I). Finally, when applied to the original histology images, several MLE predictions, based on different combinations of the three spatial statistics, could distinguish between pathologist assigned categories and the associated confidence intervals became narrower when the statistics were combined. Although the ability of $\eta_{g,F,J}$ to distinguish between manual scores is not significantly increased compared to $\eta_g$ and $\eta_{g,F}$, there is less uncertainty in evaluating $\eta_{g,F,J}$ as estimations are more precise and have narrower confidence intervals. Importantly, we do not claim that $g_{max}$, $F_{max}$ and $J_{min}$ are the optimal statistics to observe in order to best predict pathologist scores. Instead, the value of this work is as a proof-of-concept approach in which multiple statistical descriptions of a point pattern are combined to produce a single numerical description which coincides with qualitative evaluation of the point pattern.

Our training data can be expressed as a high-dimensional list of features, and therefore used to train classifiers using machine learning techniques. An advantage of using MLE over these techniques is that it is straightforward to estimate confidence intervals around $\eta$. Further, as spatial statistics describe features such as clustering or dispersion, metrics derived from their combinations can be understood in terms of the point cloud structure. Interpretation of decisions made by our approach is therefore simple compared with black-box algorithms such as neural networks where key features are often poorly understood.

The method introduced in this paper is generic in the sense that it is not tied to the particular spatial statistics employed; it can easily be extended to incorporate different spatial statistics. Possible alternatives, which have been applied to digital pathology data, include the Morisita-Horn index of colocalisation[23,24,48], Getis-Ord hot-spot analysis[26,49] and combinations of morphological characteristics[22]. Other statistics designed to characterise immune infiltration in histological data include the intratumour lymphocyte ratio[27], which measures the ratio of the number of intratumour lymphocytes and the number of tumour cells, and the Immunoscore[28–30]. Including observations of these statistics may improve predictions of pathologists' categories, and could be tested by generating synthetic validation data following our method. Importantly, adding additional statistics does not always lead to a more accurate prediction. Care must therefore be taken when selecting spatial statistical descriptors to ensure improvement. This is most likely to be obtained by using a range of statistics which identify different features of point clouds, so choosing statistics based on their function rather than making arbitrary decisions is likely to yield the best descriptors. Future work using our method will identify an optimised set of statistics which better describe immune infiltration.

Our method is not reliant on the process used to generate synthetic data; alternative processes which generate point clouds resembling immune cell distributions could be used. This flexibility means that our approach for combining spatial statistical observations could be applied more widely, given an appropriate method for generating synthetic training data. Examples of histological investigations where this may be useful include quantifying colocalisation of different immune cell subtypes, or describing relationships between immune cells and the tumour vasculature.

## Data availability

Code providing a working example of our image analysis scripts can be found at https://github.com/JABull1066/ImageAnalysisScripts. Code and data relating to combining spatial statistics and reproducing results in this manuscript can be found at https://github.com/JABull1066/CombiningSpatialStatistics.

## References

1. Saltz, J. *et al.* Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Rep.* **23**, 181–193.e7. https://doi.org/10.1016/j.celrep.2018.03.086 (2018).
2. Corredor, G. *et al.* Spatial architecture and arrangement of tumor-infiltrating lymphocytes for predicting likelihood of recurrence in early-stage non-small cell lung cancer. *Clin. Cancer Res.* **25**, 1526–1534. https://doi.org/10.1158/1078-0432.CCR-18-2013. (2019).
3. Barua, S. *et al.* Spatial interaction of tumor cells and regulatory T cells correlates with survival in non-small cell lung cancer. *Lung Cancer* **117**, 73–79. https://doi.org/10.1016/j.lungcan.2018.01.022 (2018).
4. Gartrell, R. D. *et al.* Quantitative analysis of immune infiltrates in primary melanoma. *Cancer Immunol. Res.*. https://doi.org/10.1158/2326-6066.CIR-17-0360 (2018).
5. Fridman, W. H., Zitvogel, L., Sautès-Fridman, C. & Kroemer, G. The immune contexture in cancer prognosis and treatment. *Nat. Rev. Clin. Oncol.* **14**, 717–734. https://doi.org/10.1038/nrclinonc.2017.101 (2017).
6. Heindl, A. *et al.* Relevance of spatial heterogeneity of immune infiltration for predicting risk of recurrence after endocrine therapy of ER+ breast cancer. *J. Natl. Cancer Inst.* **110**, 166–175. https://doi.org/10.1093/jnci/djx137 (2018).
7. Yuan, Z.-Y., Luo, R.-Z., Peng, R.-J., Wang, S.-S. & Xue, C. High infiltration of tumor-associated macrophages in triple-negative breast cancer is associated with a higher risk of distant metastasis. *OncoTargets Ther.* **7**, 1475–80. https://doi.org/10.2147/OTT.S61838 (2014).
8. Colling, R. *et al.* Artificial intelligence in digital pathology: A roadmap to routine use in clinical practice. *J. Pathol.* **249**, 143–150. https://doi.org/10.1002/path.5310 (2019).
9. Pell, R. *et al.* The use of digital pathology and image analysis in clinical trials. *J. Pathol. Clin. Res.* **5**, 81–90. https://doi.org/10.1002/cjp2.127 (2019).
10. Schüffler, P. J. *et al.* Automatic single cell segmentation on highly multiplexed tissue images. *Cytom. Part A* **87**, 936–942. https://doi.org/10.1002/cyto.a.22702 (2015).

11. Pham, B. *et al.* Cell counting and segmentation of immunohistochemical images in the spinal cord: Comparing deep learning and traditional approaches. *Annu Int Conf IEEE Eng Med Biol Soc.*, 842–845. https://doi.org/10.1109/EMBC.2018.8512442 (2018).

12. Cruz-Roa, A. *et al.* Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. *Sci. Rep.* **7**, 1–14. https://doi.org/10.1038/srep46450 (2017).

13. Qaiser, T. *et al.* Fast and accurate tumor segmentation of histology images using persistent homology and deep convolutional features. *Med. Image Anal.* **55**, 1–14. https://doi.org/10.1016/j.media.2019.03.014 (2019).

14. Jamaluddin, M. F. *et al.* Cell Classification in ER-stained whole slide breast cancer images using convolutional neural network. *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 632–635. https://doi.org/10.1109/EMBC.2018.8512386 (2018).

15. Laurinavicius, A. *et al.* A methodology to ensure and improve accuracy of Ki67 labelling index estimation by automated digital image analysis in breast cancer tissue. *Breast Cancer Res.* **16**, 1–13. https://doi.org/10.1186/bcr3639 (2014).

16. Nir, G. *et al.* Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts. *Med. Image Anal.* **50**, 167–180. https://doi.org/10.1016/j.media.2018.09.005 (2018).

17. Williams, B. J., Lee, J., Oien, K. A. & Treanor, D. Digital pathology access and usage in the UK: Results from a national survey on behalf of the National Cancer Research Institute's CM-Path initiative. *J. Clin. Pathol.* **71**, 463–466. https://doi.org/10.1136/jclinpath-2017-204808 (2018).

18. Schindelin, J., Rueden, C. T., Hiner, M. C. & Eliceiri, K. W. The ImageJ ecosystem: An open platform for biomedical image analysis. *Mol. Reprod. Dev.* **82**, 518–529. https://doi.org/10.1002/mrd.22489 (2015).

19. Deroulers, C. *et al.* Analyzing huge pathology images with open source software. *Diagn. Pathol.* **8**, 92. https://doi.org/10.1186/1746-1596-8-92 (2013).

20. Blom, S. *et al.* Systems pathology by multiplexed immunohistochemistry and whole-slide digital image analysis. *Sci. Rep.* **7**, 15580. https://doi.org/10.1038/s41598-017-15798-4 (2017).

21. Bankhead, P. *et al.* QuPath: Open source software for digital pathology image analysis. *Sci. Rep.* **7**, 16878. https://doi.org/10.1038/s41598-017-17204-5 (2017).

22. Wittke, C., Mayer, J. & Schweiggert, F. On the classification of prostate carcinoma with methods from spatial statistics. *IEEE Transactions on Information Technology in Biomedicine* **11**, 406–414. https://doi.org/10.1109/titb.2006.888703. (2007).

23. Yuan, Y. Spatial heterogeneity in the tumor microenvironment. *Cold Spring Harbor Perspect. Med.* **6**, a026583. https://doi.org/10.1101/cshperspect.a026583. (2016).

24. Maley, C. C., Koelble, K., Natrajan, R., Aktipis, A. & Yuan, Y. An ecological measure of immune-cancer colocalization as a prognostic factor for breast cancer. *Breast Cancer Res.* **17**, 1–13. https://doi.org/10.1186/s13058-015-0638-4 (2015).

25. Santamaría, L., Ingelmo, I., Teba, F., Rodríguez, R. & Pozuelo, J. M. Estimate of the pair correlation functions of nuclei in normal prostate, prostatic intraepithelial neoplasia, and cancer. *Eur. J. Anat.* **19**, 361–370 (2015).

26. Nawaz, S., Heindl, A., Koelble, K. & Yuan, Y. Beyond immune density: Critical role of spatial heterogeneity in estrogen receptor-negative breast cancer. *Mod. Pathol.* **28**, 766–777. https://doi.org/10.1038/modpathol.2015.37 (2015).

27. Yuan, Y. Modelling the spatial heterogeneity and molecular correlates of lymphocytic infiltration in triple-negative breast cancer. *J. R. Soc. Interface*. https://doi.org/10.1098/rsif.2014.1153 (2015).

28. Galon, J., *et al.* Type, density and location of immune cells within human colorectal tumors predict clinical outcome. *Science* **313**, 1960–1964. https://doi.org/10.1126/science.1129139 (2006).

29. Galon, J., *et al.* Towards the introduction of the 'Immunoscore' in the classification of malignant tumours. *J. Pathol.* **232**, 199–209. https://doi.org/10.1002/path.4287 (2014).

30. Pagès, F. *et al.* International validation of the consensus Immunoscore for the classification of colon cancer: A prognostic and accuracy study. *The Lancet* **391**, 2128–2139. https://doi.org/10.1016/S0140-6736(18)30789-X (2018).

31. Wang, C.-W. Robust automated tumour segmentation on histological and immunohistochemical tissue images. *PLoS ONE* **6**, e15818. https://doi.org/10.1371/journal.pone.0015818 (2011).

32. Beck, A. H. *et al.* Systematic Analysis of Breast Cancer Morphology Uncovers Stromal Features Associated with Survival. *Sci. Transl. Med.* **3**, pp. 108ra113. https://doi.org/10.1126/scitranslmed.3002564 (2011).

33. Linder, N. *et al.* Identification of tumor epithelium and stroma in tissue microarrays using texture analysis. *Diagn. Pathol.* **7**, 1–11. https://doi.org/10.1186/1746-1596-7-22 (2012).

34. Xu, J., Luo, X., Wang, G., Gilmore, H. & Madabhushi, A. A Deep Convolutional Neural Network for segmenting and classifying epithelial and stromal regions in histopathological images. *Neurocomputing* **191**, 214–223. https://doi.org/10.1016/j.neucom.2016.01.034 (2016).

35. Bulten, W. *et al.* Epithelium segmentation using deep learning in H&E-stained prostate specimens with immunohistochemistry as reference standard. *Sci. Rep.* **9**, 1–10. https://doi.org/10.1038/s41598-018-37257-4 (2019).

36. Geessink, O. G. F. *et al.* Computer aided quantification of intratumoral stroma yields an independent prognosticator in rectal cancer. *Cell. Oncol.* **42**, 331–341. https://doi.org/10.1007/s13402-019-00429-z (2019).

37. Betts, G. N. *et al.* Prospective technical validation and assessment of intra-tumour heterogeneity of a low density array hypoxia gene profile in head and neck squamous cell carcinoma. *Eur. J. Cancer* **49**, 156–165. https://doi.org/10.1016/j.ejca.2012.07.028 (2013).

38. Møller, J. & Waagepetersen, R. P. *Statistical Inference and Simulation for Spatial Point Processes* (Chapman & Hall/CRC, Boca Raton, 2004).

39. Mattfeldt, T., Eckel, S., Fleischer, F. & Schmidt, V. Statistical analysis of labelling patterns of mammary carcinoma cell nuclei on histological sections. *J. Microsc.* **235**, 106–18. https://doi.org/10.1111/j.1365-2818.2009.03187.x (2009).

40. Stoyan, D. & Stoyan, H. *Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics* (Wiley, Chichester, 1994).

41. Fozard, J. A. *et al.* Techniques for analysing pattern formation in populations of stem cells and their progeny. *BMC Bioinform.* **12**, 396. https://doi.org/10.1186/1471-2105-12-396 (2011).

42. van Lieshout, M. N. M. & Baddeley, A. A non-parametric measure of spatial interaction in point patterns. *Stat. Neerl.* **50**, 344–361. https://doi.org/10.1111/j.1467-9574.1996.tb01501.x (1996).

43. Stoyan, D. Basic Ideas of Spatial Statistics in *Statistical Physics and Spatial Statistics* (ed. Mecke, K. R. K. & Stoyan, D.) 3-21 (Springer, 2000).

44. Raue, A. *et al.* Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics* **25**, 1923–1929. https://doi.org/10.1093/bioinformatics/btp358 (2009).

45. Meeker, W. Q. & Escobar, L. A. Teaching about approximate confidence regions based on maximum likelihood estimation. *Am. Stat.* **49**, 48–53. https://doi.org/10.2307/2684811 (1995).

46. Gelman, A., Hwang, J. & Vehtari, A. Understanding predictive information criteria for Bayesian models. *Stat. Comput.* **24**, 997–1016. https://doi.org/10.1007/s11222-013-9416-2 (2014).

47. Vehtari, A., Gelman, A. & Gabry, J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27**, 1413–1432. https://doi.org/10.1007/s11222-016-9696-4 (2017).

48. Morisita, M. Measuring of the dispersion and analysis of distribution patterns. *Mem. Fac. Sci. Kyushu Univ. Ser. E Biol.* **2**, 215–235 (1959).

49. Getis, A. & Ord, J. K. The analysis of spatial association by use of distance statistics. *Geogr. Anal.* **24**, 191–206. https://doi.org/10.1111/j.1538-4632.1992.tb00261.x (1992).

### Author contributions

JAB wrote the main manuscript text and prepared figures. PSM conducted the human pathology component. TQ, FB, CWP, SLW and HMB supervised the project. All authors were involved in methodological development and investigation and reviewed the manuscript.

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41598-020-75180-9.

**Correspondence** and requests for materials should be addressed to J.A.B.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.