The Spatial Dependence between Hypoxia and Cytotoxic T Cells in Tumor Microenvironment

by

Changhao Guo
MA, University of Victoria, 2013

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Mathematics and Statistics

The Spatial Dependence between Hypoxia and Cytotoxic T Cells in Tumor Microenvironment

by

Changhao Guo
MA, University of Victoria, 2013

Supervisory Committee

Dr. Farouk Nathoo, Supervisor
(Department of Mathematics and Statistics)

Dr. Mary Lesperance, Departmental Member
(Department of Mathematics and Statistics)

Dr. Julian J. Lum, Outside Departmental Member
(Deeley Research Centre, BC Cancer and Department of Biochemistry and Microbiology)

ABSTRACT

The objective of this thesis is to examine the relationship between CAIX (a biomarker for insufficient oxygen in tumor microenvironment) and CD8+ T cells (the immune cells for killing cancer cells) for ovarian cancer. We approach the problem from two perspectives. The first approach is to set up count models such as Poisson, negative binomial, and zero-inflated Poisson models to examine the cell counts between CAIX and CD8+ T cells in the tumor microenvironment. The second approach is to apply the cross-K function, which is a second-order property of the point pattern process. We find that the tissue microarray (TMA), which is a technique to assemble hundreds of tissue samples on one TMA block, has a fixed effect on the CD8+ T cell counts. There are two TMA blocks A2 and B1. The relationship between CAIX and CD8+ T cells highly depends on TMAs. On TMA B1 stroma, a negative relationship between CAIX and CD8+ T cell counts is observed in the negative binomial models. When taking the spatial domain into account and comparing the estimated cross-K function of CAIX and CD8+ T cells to the simulated envelopes generated by a homogeneous Poisson process, we find that CAIX and CD8+ T cells are regulated and repel each other on TMA B1. Tissue category also plays an influential role in analyzing the relationship. The estimated cross-K function of CAIX and CD8 + T cells is more dispersed on tumors than on stroma.

Table of Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

First of all, I would like to express my deep and sincere gratitude to my supervisor Dr. Farouk Nathoo for supervising me to undertake the research and helping me with the completion of this thesis. I want to especially thank Farouk for taking me as a student and having patience to guide and support me. Without his help and supports, I would not be able to complete the degree and the thesis.

I'm so blessed to have Dr. Julian J. Lum as the supervisor when undertaking this research project at the Deeley Research Centre. I'm really grateful to Julian for his patience and thoughtfulness to manage the project and his constructive comments. Without this, the research would not have been possible.

I would like to thank Dr. Phineas Hamilton for his generous and precious help in training the image data and supervising me to understand pathological images and data. I'm truly thankful for his valuable assistance and suggestions.

I would like to say a big thanks to Dr. Mary Lesperance for her thoughtful comments and advice. I'm really thankful to Mary for her warmhearted encouragement after my presentations.

Many thanks to all the members at Lum lab of Deeley Research Centre and staff members at the Department of Math & Statistics for their kind support.

Last but not the least, I would like to thank my parents and sister. I'm grateful and feel extremely lucky to have their endless support and love throughout my life.

Chapter 1

Introduction

Cancer is one of the leading causes of death all over the world. Approximately 50% of Canadians are estimated to be diagnosed with cancer in their lifetime [13]. Ovarian cancer, which is a type of gynecologic cancer, has drawn much attention due to the increased number of patients diagnosed and the low survival rates. According to Ovarian Cancer Canada [1], about 1.3% of women are projected to be diagnosed with ovarian cancer. In 2020, there were 3,000 newly diagnosed ovarian cancer patients in Canada [1]. Despite the advances made in the study of ovarian cancer, survival rates have not improved significantly in recent years. The overall 5-year net survival rate in Canada is only 45% [4].

The major ovarian cancer is the epithelial ovarian carcinomas. According to the origin, development, prognosis and other features, the five major subtypes of epithelial ovarian carcinomas (EOC) are high-grade serous carcinoma (HGSC), clear cell carcinoma (CCC), endometrioid carcinoma (EC), low-grade serous carcinoma (LGSC), and mucinous carcinoma (MC) [3, 20]. HGSC is the most commonly diagnosed subtype accounting for approximate 70% of the EOC. The secondary diagnosed subtypes are CCC and EC with 10% of the EOC, respectively [3]. The remaining 10% are LGSC

and MC. When diagnosed with cancer, its severity is presented by grades that describe how the cancer cells differ from normal cells, and stages that determine the spread of cancer in the body.

Early stage ovarian cancer is asymptomatic or has mild symptoms, making it difficult to be recognized until the late stage. The statistics of the UK in 2014 shows that 55% to 58% of patients are diagnosed at stage III or IV, 42% to 45% at stage I or II, and 17% to 21% unknown stage [14]. Survival is closely related to the stage at diagnosis. The 5-year survival rate for patients in stage I is 80% to 90% in comparison to 25% for those in stage IV [19].

The high mortality rate and poor prognosis of ovarian cancer are primarily caused by late diagnosis. A tumor is a complex mass of tissue that includes cancer cells, tumor infiltrating lymphocytes cells, and stromal cells. Studies have found accumulating evidence that the outcomes of cancer are not only attributed to tumor characteristics, but also the tumor microenvironment [16]. The tumor microenvironment (TME) refers to the area surrounding a tumor cell, consisting of the blood vessels, immune cells, extracellular matrix, signaling molecules, fibroblasts, lymphocytes, and bone marrow-derived inflammatory cells. Studies have revealed that TME plays an important role in tumor growth and proliferation [64]. The characteristics of TME contain hypoxia (a condition of insufficient oxygen), acidosis (a condition of excess acid), hypoglycemia (a lack of glucose), increased cell death, and increased extracellular matrix (non-cellular components) stiffness [32, 54]. The interaction between a tumor and its surroundings is bidirectional and has a considerable impact on promoting tumor progression [64].

1.1 Relationship between Hypoxia and Infiltrating Lymphocytes Cells in the Tumor Microenvironment

On the basis of the generating mechanism, immune cells can be divided into adaptive immune cells initiated by exposure to a certain antigen and innate immune cells activated without exposure to any antigen.

Tumor infiltrating lymphocytes (TILs) are all the immune cells within tumors. As adaptive immune cells, T cells possess T-cell receptors to bind to antigens. TILs are mainly composed of anti-tumor effector T cells (cytotoxic T cells CD8+, helper CD4+ T cells) and immunosuppressive regulatory T cells (Treg). CD8+ T cells are capable of recognizing the antigen on cancer and killing cancer cells [34]. CD4+ T cells act as a helper to stimulate CD8+ T cells [36]. Large amounts of CD8+ T cells are found to be closely related to a positive prognosis in pancreatic cancer, breast cancer, colorectal cancer, lung cancer, and ovarian carcinoma [42, 43, 49, 50, 69].

Hypoxia plays an important role in tumor reproduction and has been extensively explored in the literature [21, 27]. Oxygen is an indispensable factor in metabolic pathways. Hypoxia is attributed to two sources: the lack of oxygen in blood supply or a shortage of oxygen for tumor growth.

Hypoxia areas are mainly found on solid tumors with a lower oxygen level compared to normal tissue [6]. Metabolic pathways response to hypoxia is to initiate hypoxia-inducible factors (HIFs), which can assist cancer and stromal cells to adapt the cellular microenvironment to promote the growth. HIFs contain two subunits. The α-subunit composed of HIF-1α, HIF-2α, and HIF-3α, is regulated by the oxygen levels. The

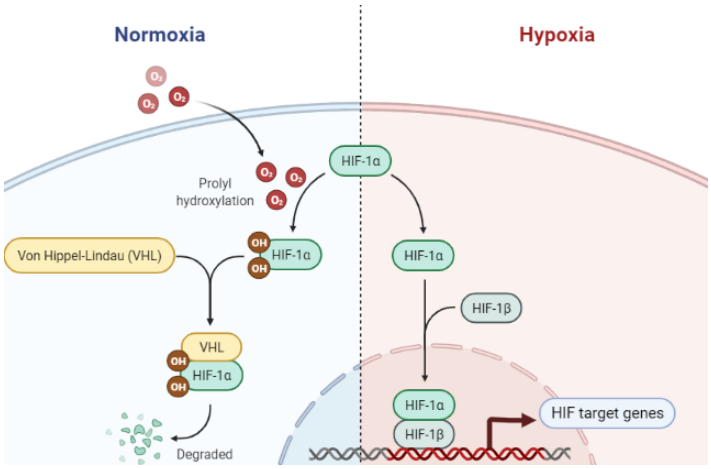


Figure 1.1: HIF pathway. In normoxia, HIF-1α is proline hydroxylated and binds the Von Hippel–Lindau protein (VHL), leading to the degradation of HIF-1α. In hypoxia, HIF-1α binds the hypoxia response elements on the target genes to promote the cell adaption to hypoxia. *Source:* Adapted from "HIF signaling", by BioRender.com (2021), retrieved from `https://app.biorender.com/illustrations/611bff68f62377009f34465`.

$\beta-$subunit comprising HIF-1β, HIF-2β and HIF-3β, is constitutively expressed [35]. Among the three family members and subunits of HIFs, the response to hypoxia is primarily regulated via HIF-1α [28]. The HIF pathway is shown in Figure 1.1. Under normoxic conditions HIF-1α is degraded, whereas under hypoxia HIF-1α is unregulated to initiate the response. Among all the target genes on HIF-1α, carbonic anhydrase IX (CAIX) plays an significant role in regulating intracellular pH for tumor growth in TME [46].

Hypoxia results in resistant to the chemotherapy and is strongly associated with clinical treatment outcomes and patient survival [61, 63]. The studies have discovered that CAIX is closely related to poor prognosis in cancers such as lung [59], breast [17], head and neck [30]. Regarding ovarian cancer, CAIX is usually found to be overexpressed with magnitudes of the expression varying with subtypes. CAIX is highly associated with low survival and poor prognosis, and therefore has been proposed as an adverse prognostic marker for EOC [18].

In regard to the effects of hypoxia on T cells, the statement is still ambiguous and under investigation. Under the condition of deficient oxygen, the expression of HIF-1α increases. HIF-1α impacts T cells through depressing the activation, promotion, and ability to kill cancer cells [47, 51]. The deletion of HIF-1α gene reinforces the T cell responses [60]. Barsoum's study of breast cancer, prostatic carcinoma, and murine cancer found that hypoxia promoted cancer cells to escape from cytotoxic T cell response through the HIF-1α expression of programmed cell death ligand-1 (PD-L1) [10]. In the meanwhile, there is evidence demonstrating that CD8+ is inclined to differentiate more efficiently into cytotoxic T cells under deficient oxygen [9]. On the other hand, T cells respond to HIF-1α by evading the hypoxia area. Most T cells are discovered in the non-hypoxic areas.

1.2 Research Objective

Regarding ovarian cancer, HIF expression by hypoxia is frequently found in ovarian cancer patients [31, 66]. The HIF stabilization is closely correlated with the co-expression of tumor-infiltrating Treg cells CD8+ and FoxP3 [25]. The genetic expression of hypoxia is also detected [68].

Nevertheless, the relationship between hypoxia and T cells is not well investigated. In this thesis, we examine this relationship by examining the relation between the biomarkers CAIX and CD8+ T cells in tumor images from the perspective of the tissue level and the cellular level data, respectively. It may have important implications and offer opportunities to reveal the interaction within TME for potential cancer therapies.

Our analysis considers summary statistics, regression analysis, and spatial statistics. The commonly used spatial analysis of cell data in the literature are the density and

the nearest neighbors. For example, Zheng et al. [71] investigated the density and the spatial distribution of tumor-associated macrophagesarer in lung cancers, Masugi et al. [42] revealed the importance of spatial distribution of CD8+ T cells for pancreatic cancer prognosis. The cross-type nearest neighbor distance function (Cross-G) was also implemented to account for the spatial relationships between tumor and T cells [11]. But a serious defect of the nearest neighbors is ignoring the other spatial information about cell locations, except the nearest cells. In this study, we apply the spatial statistics cross-K function to account for all the spatial information of cell locations.

1.3 Organization of Thesis

The organization of the thesis is as follows. The first chapter provides some background and descriptions of tumor hypoxia and T cells in the tumor microenvironment.

In Chapter 2, we present a review of some of the methods applied in our analysis. Resting on the data collection and how we define an observation, we undertake analysis with two approaches. Cells form tissues, and on the condition that each tissue culture is an observation, the tissue data is summarized by the number of cells on a tissue sample as count data. Therefore, regression models based on the Poisson model, zero-inflated Poisson model, and the negative binomial distribution are employed and the relative fit computed. An alternative approach is to view individual cells as the realization at a point pattern and implement a spatial analysis. The techniques are discussed in Chapter 2.

Chapter 3 is a case study of ovarian cancer. We explore the effect of hypoxia on T cells using the two approaches described in Chapter 2. Data on the tissue level are modeled and analyzed from the perspective of correlation statistic, negative binomial

models, and a zero-inflated model. Moreover, cross-K functions are used for the cellular data to investigate the spatial structure of the TME. The last chapter consists of conclusions and a discussion of future work.

Chapter 2

Theoretical Background

This chapter covers the methodologies we employed. The data on the tissue level is count data, which are non-negative integers to represent the number of occurrences of an event in a spatial domain. Typical examples of count data are the number of patients, the number of cells, and the number of doctor visits.

An initial step to investigate the linear interaction between two variables is Pearson correlation statistic. Since we are interested in the impacts of hypoxia on T cells, a model to regress T cells on hypoxia measures can explore the magnitude of its association. By taking cell coordinates into account, the data on the cellular level involves more spatial information. Spatial point pattern analysis is used to investigate the spatial structure between cell point locations.

2.1 Pearson and Spearman Correlation Coefficients on the Tissue Level

To measure the correlation between two variables, the simplest and most commonly used approach is the Pearson correlation coefficient to evaluate and estimate the linear

relationship. The formula is given by:

$$\rho = \frac{\sum_{i=1}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}(x_i - \bar{x})^2 \sum_{i=1}(y_i - \bar{y})^2}} \tag{2.1}$$

The range of ρ value is from -1 to 1. A value of ρ close to zero implies little correction. A positive ρ indicates a positive linear relationship, whereas a negative value indicates a negative relationship.

The Spearman correlation coefficient is a non-parametric version of the Pearson correlation coefficient to measure the monotonic relationship between two variables. It can be computed using the formula:

$$\rho_s = 1 - \frac{6 \sum_{i=1} D_i^2}{n(n^2 - 1)} \tag{2.2}$$

where D_i is the difference between the observation ranks and n is the number of observations. ρ_s has the same interpretation as ρ but allows for non-linear relationship.

2.2 Regression Models on the Tissue Level

In this section, we present descriptions of the Poisson, negative binomial, and zero-inflated regression models subsequently used in our analysis.

2.2.1 Generalized Linear Models

When the response variable for regression analysis is count data, the classical linear regression model will not be suitable due to the violation of the normality assumption. To allow for the non-normality of the data, the linear regression model can be extended to a generalized linear model (GLM), which allows generally for response variables with

the exponential family [45]. There are three primary components in a GLM: a random component that specifies the distribution of the response variable, a linear predictor that is a linear combination of the covariates and parameters, and a link function that connects the linear predictor to the mean of the response variable. A fourth component that is important in some cases is a function that relates the mean and the variables. The linear regression model is a special case of a GLM on the condition that the response variable is normally distributed and an identity link is applied.

Given the response variable $y_i, (i = 1, 2, ..., n)$ and a link function $g(.)$, the regression aspect of GLM is represented by:

$$\eta_i = g(\mu_i) = g[E(y_i)] = \boldsymbol{x_i^T \beta} \tag{2.3}$$

where $\boldsymbol{x_i}$ is a vector of covariates for ith observation and $\boldsymbol{\beta}$ is a vector of parameters. The distribution of the response variable is now not limited to normality or linearity with respect to $E(y_i)$. Distributions from the exponential family, including the binomial, Poisson, or negative binomial can be employed.

The link function is a monotonic and differentiable function that can take many forms. The commonly used links are the identity link ($\eta_i = \mu_i$) for a normal distribution, the logistic link ($\eta_i = \ln(\frac{\pi_i}{1-\pi_i})$) for a binomial distribution, the log link ($\eta_i = \ln(\mu_i)$) for a Poisson distribution, and the reciprocal link ($\eta_i = \dfrac{1}{\mu_i}$) for a gamma distribution.

Parameter Estimation

The maximum likelihood estimation (MLE) determines estimates for parameters by maximizing the log-likelihood function of the data. Let the observations $\boldsymbol{y} = (y_1, ..., y_n)$

with the probability density function $f(y_i;\boldsymbol{\beta})$ and parameters $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$, the log likelihood function of the data is:

$$l(\boldsymbol{\beta};\boldsymbol{y}) = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} \ln\ f(y_i;\boldsymbol{\beta}) \tag{2.4}$$

To obtain the MLE, we take the first order conditions of l with respect to each individual β to maximize this log-likelihood function. The derived equations are referred to as score equations.

$$\left(\frac{\partial l(\boldsymbol{\beta};\boldsymbol{y})}{\partial \beta_1}, ..., \frac{\partial l(\boldsymbol{\beta};\boldsymbol{y})}{\partial \beta_p}\right)^T = \mathbf{0} \tag{2.5}$$

The iteratively reweighted least squares (IRLS) is used to solve the score equations, which is an iterative process to compute the MLE using weighted least squares. We first choose $\boldsymbol{b_0}$ as an initial estimate of $\boldsymbol{\beta}$. Then for $k = 0, 1, 2, \ldots$ iteration, we compute the estimated mean $u_i^{(k)} = g^{-1}(\boldsymbol{x_i^T b^{(k)}})$ and the adjusted response variable $Z_i^{(k)} = \boldsymbol{x_i^T b^{(k)}} + g'(u_i^{(k)})(y_i - u_i^{(k)})$. We subsequently derive the weights $W^{(k)} = W(b^{(k)})$. The final step is to regress $Z^{(k)}$ on the design matrix X with weight $W^{(k)}$ to obtain a new estimate $\boldsymbol{b}$. This iteration process continues until convergence. The converged values are the MLE of $\boldsymbol{\beta}$.

Goodness of Fit

We can use the likelihood ratio test to compare a full model with a reduced model. Let the log likelihood function for the full model be l_f and the log likelihood function of the reduced model (for example no covariates) be l_r, the test statistic is

$$LR = 2(l_f - l_r) \tag{2.6}$$

LR approximately follows a χ^2 distribution with a degree of freedom equal to the number of parameters of the full minus the number of parameters of the reduced model. The null hypothesis is that the reduced model is preferred. A large value of LR (i.e., small p-value) indicates a rejection of the null and the full model is an adequate fit.

Deviance can also be used to check the goodness of fit. Let the log likelihood function for the saturated model be l_s, where each observation has its own parameters. The deviance is defined as:

$$D = 2(l_s - l_f) \tag{2.7}$$

D approximately follows a χ^2 distribution with degree of freedom equal to $(n - p)$, where p is the number of parameters in the full model. A small deviance (or a large p-value) implies that the model is an adequate fit. A rule of thumb is that if $\frac{D}{n-p}$ is substantially greater than 1, the model is inadequate [48].

The Pearson χ^2 test statistic is also an alternative approach to test the goodness of fit, which is defined as:

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{y_i - \hat{\mu}_i}{\sqrt{\hat{Var}(y_i)}} \right)^2 \tag{2.8}$$

The Pearson statistic also follows a χ^2 distribution with $(n - p)$ degree of freedom. A small value indicates that the model is adequate.

Residual Analysis

Residual analysis can be used to examine the model assumptions. The most frequently used is the Pearson residual $r_p = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{Var}(y_i)}}$. The alternative is a deviance residual, which is defined as $d_{i,r} = [sgn(y_i - \hat{\mu}_i)] \times \sqrt{d_i}$. The sign of the deviance depends on $(y_i - \hat{\mu}_i)$ and it also satisfies that $D(\beta) = \sum_{i=1}^{n} d_{i,r}^2$. We can find how these residuals

change by plotting these residuals against the fitted values or covariates.

2.2.2 Poisson Model

When the response variable $y_i (i = 1, 2, ...n)$ is count data, a commonly used probability model is the Poisson distribution. The probability mass function of a Poisson distribution is:

$$f(y_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} \tag{2.9}$$

where μ_i is the parameter for the Poisson. The mean and variance for Poisson are both equal to parameter μ_i. In regard to the link function, a popular choice is the log link $\eta_i = \ln(\mu_i) = \boldsymbol{x_i^T \beta}$, which can ensure the predicted response variable is non-negative. Thus, the relation between the mean of the response variable and the linear predictor can be captured by:

$$\mu_i = e^{\boldsymbol{x_i^T \beta}} \tag{2.10}$$

The Poisson regression has been widely used in assessing the cell counts. For example, Lodwick et al. [39] used Poisson regression to analyze the relation between mortality rate and CD4 cell counts for HIV patients.

When using the Poisson regression, one of the key assumptions is that the mean and variance are the same $E(y_i) = Var(y_i) = e^{\boldsymbol{x_i^T \beta}}$. However, this is not always true in practice. The variance of observations may be larger (or smaller) than the mean value, making Poisson regression inappropriate to use. When the variance is larger than the mean value, this phenomenon is referred to as overdispersion [48], whereas the variance being smaller than the mean is said to be underdispersion [48]. Compared to

underdispersion, overdispersion is more frequent in empirical data, such as the vaccine data [57], cigarette and marijuana usage [53], and biomarker counts [15].

Overdispersion can arise from heterogeneity, dependency within observations, the presence of outliers, and the existence of excess zeros [48]. Given the presence of overdispersion, Poisson regression tends to underestimate the standard error, narrow down the confidence interval of the estimates, and therefore lead to incorrect under-conservative statistical inferences and conclusions [29].

Depending on the sources of overdispersion and the mechanisms for tackling this problem, a number of models are proposed to account for overdispersion in the literature. An alternative model that allows for overdispersion is the negative binomial model, where the variance can be larger than the mean. When the overdispersion is caused by abundant zeros in observations, the zero-inflated Poisson model is also an alternative model though it is not an exponential family.

2.2.3 Negative Binomial Model

On the basis of the Poisson log-linear model, an alternative approach to allow for overdispersion in count data is the negative binomial model that incorporates an additional parameter to model the excess variance [52].

The negative binomial model is a Poisson-gamma mixed distribution, where the response variable follows a Poisson distribution $Pois(\mu)$ conditional on the random parameter μ. Integrating out μ from the joint distribution then yields the negative binomial. In this case, the negative binomial can accommodate more variation in the response variable. The probability mass function is:

$$f(y_i;\mu,\alpha) = \frac{\Gamma(y_i + 1/\alpha)}{y_i!\Gamma(1/\alpha)}(\frac{1}{1+\alpha\mu_i})^{1/\alpha}(\frac{\alpha\mu_i}{1+\alpha\mu_i})^{y_i} \tag{2.11}$$

with mean $E(y_i) = \mu$ and variance $V(y_i) = \mu + \alpha\mu^2$, and $\alpha \geq 0$ is defined as the dispersion parameter. The Poisson distribution is a special case of the negative binomial arising when $\alpha = 0$. As α increases, the variance is greater than the mean by adding a quadratic term. This is how the negative binomial captures the overdispersion, though other forms of overdispersion could be considered.

Similar to a Poisson regression, the negative binomial regression is a GLM. Given a log link function $g(\mu_i) = \ln(\mu_i) = \boldsymbol{x_i^T\beta}$, the mean of response variable is a function of linear predictor $\mu_i = e^{\boldsymbol{x_i^T\beta}}$.

2.2.4 Zero-inflated Poisson Model

In practice, overdispersion may be caused by an excess of zeros in observations. For example Rodrıguez [56] found that although the Poisson model predicted that 21% of biochemistry PhD graduates did not get papers published, the actual data showed this proportion was 30% .This 9% gap was primarily a result of the Poisson model being a poor fit that could not predict the number of zeros in the data. When analyzing the tumor immune microenvironment, many tissue samples had zero positive cells that need to be taken into account [65].

An alternative model to tackle the excess zeros is the zero-inflated Poisson model, which assumes that the distribution of observations is a mixture of two distinct distributions. Hence, the zero observations have two sources: the structural zeros from non-affected groups that yield zeros with probability one, and the non-structural zeros from affected groups whose probability at zero depends on some covariates and the Poisson distribution [33]. The zero-inflated model allows for estimation of these two distributions or each group, respectively.

A zero-inflated Poisson model (ZIP) is proposed by Lambert [38], where the ob-

servations are generated from two possible distributions: a point mass at zero with a probability p_i and a Poisson model $Pois(\lambda_i)$ with a probability $(1-p_i)$. The probability mass function of a ZIP model is:

$$Pr(y_i) = \begin{cases} p_i + (1-p_i)e^{-\lambda_i} & \text{for } y_i = 0 \\ (1-p_i)\frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} & \text{for } y_i > 0 \end{cases} \tag{2.12}$$

where the mean is $E(y_i) = \mu_i = (1-p_i)\lambda_i$ and variance is $V(y_i) = \mu_i + \frac{p_i}{1-p_i}\mu_i^2$, respectively. As p_i approaches zero, the mean and variance get closer and the ZIP model reduces to the Poisson model. In addition, when the probability p_i is expressed with a logit link function and λ_i is expressed with a log link function, the regression model is:

$$logit(p_i) = \ln(\frac{p_i}{1-p_i}) = \mathbf{z_i^T}\gamma \tag{2.13}$$

$$\ln(\lambda_i) = \mathbf{x_i^T}\beta \tag{2.14}$$

where $\mathbf{z_i}$ is a vector of covariates and γ is a vector of regression coefficients for the zero-inflation part, respectively. $\mathbf{x_i}$ and β are the covariates and regression coefficients for the ZIP model. Depending on the choice of covariates, $\mathbf{z_i}$ and $\mathbf{x_i}$ may or may not be the same.

2.2.5 Comparison of the Negative Binomial and Zero-inflated Poisson Model

We consider three count models in this chapter: Poisson, negative binomial, and the ZIP models. The Wald test and likelihood ratio test are only capable of comparing nested models. The Poisson is nested within the negative binomial model, but the neg-

ative binomial and ZIP model are not nested. For non-nested models using maximum likelihood methods, the information criteria can be employed to select models.

Overdispersion Test

Since the Poisson is nested within the NB model, we can implement a likelihood ratio test to test the Poisson model against overdispersion [48]. Given the dispersion parameter α, the null hypothesis is on the boundary of parameter space $H_0 : \alpha = 0$, whereas the alternative hypothesis is $H_A : \alpha > 0$. The likelihood ratio test statistic is:

$$LR = -2[l(\hat{\mu}) - l(\hat{\mu}, \hat{\alpha})] \tag{2.15}$$

where $l(\hat{\mu})$ and $l(\hat{\mu}, \hat{\alpha})$ are the maximized log likelihood under the Poisson and the NB regressions. Under the null hypothesis, LR is asymptotically χ^2 distributed with 1 degree of freedom.

Information Criterion

Akaike Information Criterion (AIC) and Schwarz Bayesian Information Criterion (BIC) have been frequently used for model selection. The idea of AIC is to assess the goodness of fit and to penalize by adding the number of parameters to the model [5]. The formula of AIC is:

$$AIC = -2L + 2k \tag{2.16}$$

where L is the maximized log-likelihood and k is the number of estimated parameters.

BIC follows the same idea but with a different penalty term. BIC penalizes stronger

than AIC [44]. With the number of observations n, the formula of BIC is:

$$BIC = -2L + k\ln(n) \tag{2.17}$$

The AIC and BIC for Poisson, negative binomial, and ZIP will be computed under maximum likelihood estimation and the model with minimum AIC or BIC is preferred.

2.3 Point Pattern Analysis on the Cellular Level

2.3.1 Point Pattern and Point Process Models

Point pattern data are a set of locations of objects (or events) in a specified study region that are considered the realization of a random process [7]. The study region can be two-dimensional, three-dimensional, or multi-dimensional. In a two-dimensional region, the location is typically recorded with Cartesian coordinates (x, y), which is defined by x units on the horizontal axis and y units on the vertical axis. Each location is thus represented by a point. In addition, if there are other data (or attributes) attached to the points, they are referred to as marks. Points with marks are called marked points and result in what is known as a marked point process.

A point process is a random process that generates the distribution of points in a study region. Each point represents the time of an event in a temporal process. In contrast each point represents the location of an event in a spatial process. Spatio-temporal processes constitute both time and location (x, y, t). In our study, the sample data are two-dimensional spatial data representing biomarker locations in a tumor image.

Point pattern analysis is about the understanding of a point process. Determining

the type of point process model (such as Poisson or Cox process) and examining the spatial clustering/regularity between points (independent or dependent) are typical research questions in this field.

The approaches can be divided into first-order properties and second-order properties [26]. The first-order properties examine the expected number of points in a study region, whereas the second-order properties focus on the correlation between points. The first-order properties are characterized by the intensity function and its kernel estimation. As regards to the second-order properties, the K function provides a possible method to explore the data.

2.3.2 Intensity

Denoted as λ, intensity is the expected number of points per unit area. The mathematical definition of intensity is [23]:

$$\lambda(s) = \lim_{ds \to 0} \left\{ \frac{E[N(ds)]}{ds} \right\} \tag{2.18}$$

where ds is a small region around the point and $N(ds)$ is the number of points observed in this region ds. The intensity can be constant or vary across regions. A constant intensity means that the intensity is the same anywhere in the study region, which is referred to as a homogeneous process. Given the number of points $n(A)$ in a study region A, the intensity for a homogeneous process is simply $\frac{n(A)}{|A|}$. When the intensity varies from location to location, this is referred to as an inhomogeneous process, which can model non-stationary data. One approach to estimate the intensity is the quadrat method. The idea is to partition the study region into small sub-regions (or quadrats) and to count numbers of points in each quadrat. Another approach is kernel smoothing,

which produces an estimate $\lambda(s)$ at each point s weighting points within a neighborhood of s according to a kernel function [26]. In our study, we assume that the intensity in a study region is constant and the data arise from a homogeneous point process.

2.3.3 Complete Spatial Randomness

The simplest point process is the spatial homogeneous Poisson process, under which the number of points in a region is Poisson-distributed. Spatial homogeneous Poisson process is also referred to as Complete Spatial Randomness (CSR). Under CSR, all the points are independently and uniformly distributed in a study region and the intensity or the mean of points per unit area is assumed constant [23]. Even though CSR is hardly seen in practice, it is often used as a null hypothesis to investigate whether points are random, clustered, or regulated.

Hypothesis testing of CSR can be constructed from either the nearest neighbor or second-moment properties [22].

2.3.4 The F function and G function

The G function measures the distribution of distance from a point to its nearest neighbors [23]. Given there are n points in the study region, the empirical distribution function $\hat{G}(d)$ is:

$$\hat{G}(d) = \frac{1}{n}\#(d_i \leq d) \tag{2.19}$$

where d_i is the distance from the ith point to its nearest neighbors and $\#(.)$ denotes the number of points in the given region. Under CSR with a constant intensity λ, the

explicit theoretical expression of $G(d)$ is:

$$G(d) = 1 - e^{\lambda \pi d^2} \tag{2.20}$$

The F function measures the probability of an arbitrary point having a point in the neighborhood within a distance d [23]. Given there are m number of arbitrary points selected to find the nearest point, the empirical distribution function $\hat{F}(d)$ is:

$$\hat{F}(d) = \frac{1}{m} \#(d_i \leq d) \tag{2.21}$$

Under CSR, the theoretical expression of $F(d)$ is identical to $G(d)$.

By comparing the empirical function $\hat{G}(d)$ (or $\hat{F}(d)$) with the theoretical form for $G(d)$ (or $F(d)$) under CSR, we are able to identify the degree of departure form CSR and thus if the underlying points in the study region exhibit clustering (or a regular) pattern.

2.3.5 The K Function

One of the drawbacks of the F function and G function is that only the nearest neighbors are incorporated. The information about distance is incomplete and the influence of edge effects is not taken into account [22]. As an alternative, we can use (Ripley's) K-function, which is a second-order method to allow for all the pairwise distances.

A spatial point process is isotropic if it is the same in any direction. For a stationary isotropic point pattern, the K function is defined as [24]:

$$K(d) = \frac{E(\text{number of points within radius } d \text{ of an arbitrary point})}{\lambda} \tag{2.22}$$

where λ is the intensity of the point process.

In a homogeneous Poisson process (or CSR), the intensity λ is a constant number $\bar{\lambda}$. The expression for the K function is simply given by:

$$K(d) = \frac{\bar{\lambda} \cdot \pi d^2}{\bar{\lambda}} = \pi d^2 \tag{2.23}$$

$K(d) \geq \pi d^2$ indicates clustering, since there are more points than expected under the theoretical form for CSR. In contrast, $K(d) \leq \pi d^2$ indicates a regular pattern, since there are less points than expected.

Various estimators of K have been proposed in the literature. The most commonly used estimator with edge effects is given by:

$$\hat{K}(d) = \frac{1}{\hat{\lambda}^2 A} \sum_i \sum_{j \neq i} \omega_{ij}^{-1} \, I(d_{ij} < d) \tag{2.24}$$

where A is the area of the study region, $\hat{\lambda}$ is the estimated intensity, d_{ij} is the distance between ith and jth point, and $I(.)$ denotes the indicator function. The weight function ω_{ij} provides an edge correction.

Similar to the G function (or F function), we can plot the empirical function $\hat{K}(d)$ against the theoretical form $K(d)$ under CSR. For a homogeneous Poisson process, spatial clustering is indicated by $\hat{K}(d) > \pi d^2$, while a regular pattern is represented by $\hat{K}(d) < \pi d^2$.

2.3.6 The Cross-K Function

The K function only takes the location of an event into account, for example cell locations in a tissue, the patient locations in a city, or the tree locations in a region.

But each point may carry additional information, such as the phenotype of each cell, the gender of the patient, or the species of the tree. The additional information is presented as the mark for the point. A point pattern with more than one mark is referred to as a multivariate point pattern. When working with a multivariate point pattern, the previous K function yields to one univariate process; whereas we have a multivariate process. However, in practice, we are also interested in the relationships between different types of points. This gives rise to a generalization of the K function for more than one type of point to allow for interactions. An example is to investigate the relationship between CD8+ T cells and regulatory T cells. The cross-K function of point i in respect to point j is given by [24]:

$$K_{ij}(d) = \frac{E(\text{number of type } j \text{ point within a radius } d \text{ of a type } i \text{ point})}{\lambda_j} \tag{2.25}$$

The Ripley's estimator of the cross-K function is similar to that of the univariate K function:

$$\hat{K_{ij}}(d) = \frac{1}{\hat{\lambda_i}\hat{\lambda_j}A} \sum_k \sum_l \omega_{i_k,j_l}\, I(d_{i_k,j_l} < d) \tag{2.26}$$

where A is the area of study region, $\hat{\lambda_i}$, and $\hat{\lambda_j}$ are estimated intensity, d_{i_k,j_l} is the distance obtained from the data between the kth location of type i and the lth location of point type j, $I(.)$ denotes the indicator function; and ω_{i_k,j_l} is the edge correction function. K_{ij} is the same as K_{ji}. Moreover, K_{ii} is just the univariate K function K_i. The examination of the cross-K function can reveal potential relationship between two point processes.

Suppose there are two types of points with subscript $i = 1$ and $j = 2$. Under the null hypothesis that the two types of points are generated by an independent process, the cross-K function is represented by $K_{12} = \pi d^2$. When $\hat{K}_{12} > \pi d^2$, it indicates that

the two types of points tend to cluster, that is to show up together. When $\hat{K}_{12} < \pi d^2$, it suggests that the two types of points tend to be regulated and dispersed.

Diggle (2013) [23] indicated cross-K function and other summary functions (K, F and G functions) should be used together with Monte Carlo simulation envelopes to provide some statistical variability. Suppose that there are K times independent simulations of cross-K function under CSR. For a given distance, the maximum value of the simulated cross-K function becomes the upper bound; whereas the minimum value of the simulated cross-K function formulates the lower bound. The interval between lower and upper bound is the simulated envelope. If the empirical function falls outside the simulation envelope bounds, this provides evidence against CSR. If the empirical function falls above the upper envelope, more points are observed at the given distance and the point pattern is clustered. If it lies under the lower bound, less points are observed and this implies points are inhibited [26]. Figure 2.1 illustrates the estimated cross-K function of two biomarkers. The shadowed area is the simulated envelope. We can see that the estimated cross-K function lies outside the envelope, implying a deviation from CSR.

2.4 Concluding Remarks

Depending on the way the data are collected, there are two approaches to analyze the interaction between cells within the tumor microenvironment (TME). We enumerate the theoretical models for tissue level data and cellular level data, respectively. The explicit models for our study will be presented in the next chapter.

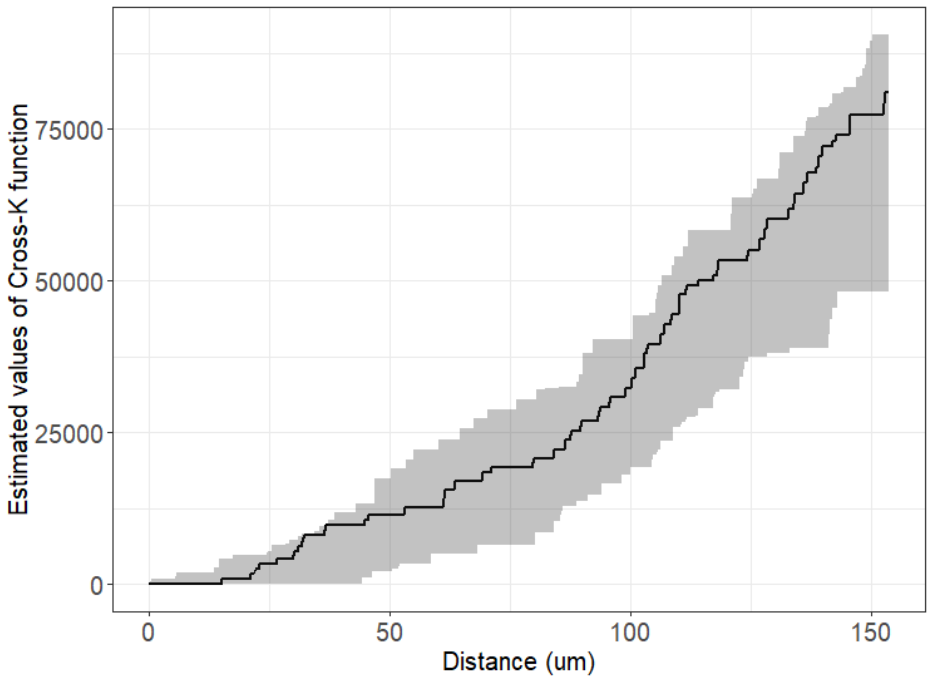


Figure 2.1: The estimated cross-K function of CAIX and CD8+ cells for COEUR TMA block A2-core A-14 with an envelope derived from 99 Monte Carlo simulations.

Chapter 3

Analysis of Ovarian Cancer

Hypoxia plays an important role in the development of tumors. Increasing evidence indicates that hypoxia is significantly associated with poor prognosis and unsatisfying therapy outcomes [61, 63]. As a result, exploring the impacts of hypoxia on ovarian cancer will enhance our understanding of the effects of hypoxia and promote the development of cancer therapies.

The effects of hypoxia on ovarian cancer have been discussed in the literature, including investigating the target gene expression under hypoxia [70], detecting the molecular mechanism on hypoxia regulation [68], enhancing the ovarian cancer cell responses [37], and the existence and functions of the hypoxia-inducible factors (HIFs) [55]. In this study, our objective is to examine the relationship between hypoxia and the distribution of T cells in the TME.

Carbonic Anhydrase IX (CAIX) is an enzyme to catalyze the reversible transformation from carbon dioxide and water into carbonic acid, proton, and bicarbonate anion ($H_2O + CO_2 = H^+ + HCO_3^-$) [40] and is regulated by the HIF. Among the two subunits, the β-subunit is constantly expressed and insensitive to changes in oxygen levels. But the expression of the other subunit HIF-1α heavily depends on the oxygen

conditions. As we discussed in Chapter 1, under hypoxic conditions, the prolyl hydroxylase domain-containing enzymes cannot be produced without sufficient oxygen. Away from degradation, HIF-1α becomes stabilized and generates heterodimer with HIF-1β. The heterodimer binds hypoxia response elements of the target gene and promotes the adaptation of cells to a hypoxic microenvironment [70].

CAIX has been used by researchers as a biomarker for hypoxia in tumors. Evidence shows that there is a high level of HIF-1α in ovarian cancer [55]. Therefore, we use CAIX as the biomarker for hypoxia in the ovarian cancer study.

The immune system is initiated under the emergence of a tumor. The naive CD8+ cells, a T cell that differentiates in the thymus, are first differentiated into effector CD8+ T cells. With the interaction between TCR and the antigenic peptide–major histocompatibility complex (MHC), the effector CD8+ T cells are activated into cytotoxic and memory CD8+ T cells to fight against tumor cells [41]. These targeting functions of CD8+ T cells are undertaken by producing cytolytic enzymes such as granzyme B, interacting with receptors on cancer cells, and secreting cytokines such as TNFα to induce the death of cancer cells [2].

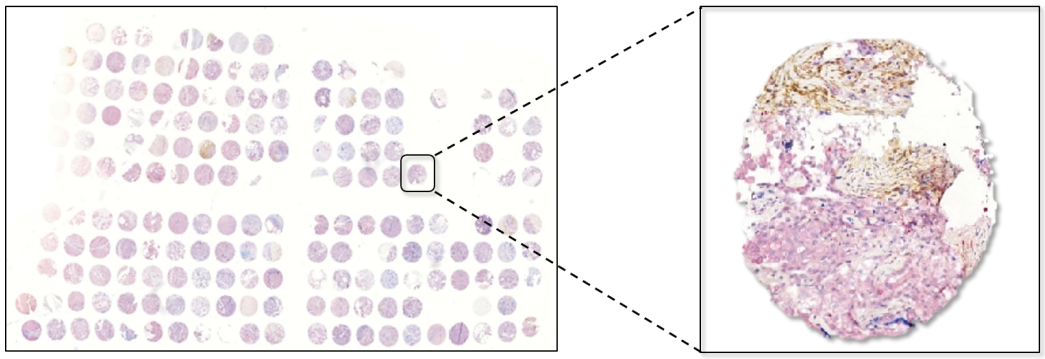
The research question is focusing on investigating the relationship between hypoxia and immune cells. To be more specific in our study, we are examining the relationship between CAIX and CD8+ T cells in tumor images. Applying the methods discussed in Chapter 2, the analysis is conducted in two levels: the tissue level and the cellular level. The explicit models are built in the methods section.

3.1 The COEUR Cohort

We concentrate on biological data from clear cell carcinoma, one of the five histological subtypes in ovarian cancer. The tissue samples of clear cell carcinoma ovarian cancer were obtained from the Canadian Ovarian Experimental Unified Resource (COEUR), a platform to collect human epithelial ovarian cancer biological materials. The tissue samples were obtained, cleaned and arrayed on tissue microarrays (TMAs). Tissue microarrays (TMAs) are technologies to allow for simultaneously analyzing many tumor samples on a single microscope glass slide, where hundreds of cylindrical holes are punched and arrayed [58]. A sector map is first constructed as a reference for assembling and scoring the TMA cores. A typical diameter of a cylindrical tissue core is 0.6 mm and a representative tumor area from a donor is put in the core according to the sector map.

Additionally, TMAs are hematoxylin & eosin (H & E) stained to detect cellular and tissue structure. We make use of 2 biomarkers staining to identify the CAIX and CD8+ T cells. After TMA slides are digitally scanned into a computer, the next step is to score these digital images. The TMA scoring process is performed in *QuPath* [8] and manually assessed by an expert.

The 202 donor patients in the COEUR cohort assemble two TMA slides (A2 and B1). The scored cells are cataloged into three phenotypes (CAIX, CD8, and others), and two tissue structures (tumor and stroma). TMA A2 assembles 173 valid cores and TMA B1 has 203 valid cores. Figure 3.1 illustrates the COEUR TMA slide B1 and its corresponding sector map. The brown staining area is CAIX and the dark blue stains are CD8+ T cells. It is not unusual to see that the number of valid cores on the TMA B1 is less than the number of cores on the sector map, as 19 cores are dropped due to



(a)

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

101	102	103	104		105	106	107	108	109
110	111	112	113		114	115	116	117	118
119	120	121	122			123	124	125	126
127	128	129	130			131	132	133	134
135	136	137	138			139	140	141	142
143	144	145	146	147	148	149	150	151	152

51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
70	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

153	154	155	156	157	158	159	160	161	162
163	164	165	166	167	168	169	170	171	172
173	174	175	176	177	178	179	180	181	182
183	184	185	186	187	188	189	190	191	192
193	194	195	196	197	198	199	200	201	202
203	204	205	206	207	208	209	210	211	212
213	214	215	216	217	218	219	220	221	222

(b)

Figure 3.1: COEUR TMA slide B1 and its TMA map. (a) A hematoxylin & eosin stained TMA slide B1. The brown stains are hypoxia area and the blue stains are CD8+ T cells. (b) The sector map of TMA layout containing 222 cores.

unclear scanning or no scored cells.

3.2 Methods

The Negative Binomial and Zero-inflated Poisson Regression Models

We will start with the tissue level analysis where the variables are cell counts of different phenotypes computed at each tumor image. Given that there are S tissue samples on a TMA slide ($S = 173$ for TMA A2 and $S = 203$ for TMA B1), CAIX cell counts x_i $(i = 1, 2, \ldots, S)$ and CD8+ T cell counts y_i $(i = 1, 2, \ldots, S)$ are used to represent the accumulated cell counts for each tissue sample. Since the image cores are nested within TMA slides, there may exist a fixed effect (FE) for these blocks. Therefore, we incorporate a fixed effect represented by z_i in the statistical model. There are two

Table 3.1: Count models for CD8+ T cell counts y_i with coefficents β, covariates CAIX cell counts x_i, TMA fixed effect z_i, and the point mass probability p_i in ZIP model.

Models	Model specification
Poisson	$ln(E(y_i)) = \beta_0 + \beta_1 x_i$
NB	$ln(E(y_i)) = \beta_0 + \beta_1 x_i$
NB with fixed effect (FE)	$ln(E(y_i)) = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i$
ZIP	$ln(E(y_i)) = \beta_0 + \beta_1 x_i$ with a probability $(1 - p_i) = \frac{1}{e^{\gamma_0}+1}$
ZIP with fixed effect (FE)	$ln(E(y_i)) = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i$ with a probability $(1 - p_i) = \frac{1}{e^{\gamma_0}+1}$

TMA slides A2 and B1 in our study cohort. $z_i = 0$ if the tissue sample lies on TMA A2 and $z_i = 1$ if the tissue sample is from TMA B1.

Given the response variable of CD8+ T cell counts y_i and the CAIX cell counts x_i, fixed effects z_i, and the interaction between x_i and z_i, Poisson, ZIP, and NB regression models are considered. Table 3.1 lists the five models considered. The log of CD8+ T cells is regressed on an intercept and a covariate in the Poisson and NB regression models. The NB regression model allows for the variance being greater than the mean. The NB with fixed effect regression model incorporates additional covariate z_i and the interaction between x_i and z_i. In regard to the ZIP models, only an intercept is included in the regression of (2.12). That is, the probability of y_i generated from a point mass at zero is $p_i = \frac{e^{\gamma_0}}{e^{\gamma_0}+1}$ and with a probability of $(1 - p_i) = \frac{1}{e^{\gamma_0}+1}$, y_i is generated from a Poisson distribution. The Poisson regression part of (2.13) in the ZIP regression model is to regress the log of CD8+ T cells on the covariates.

The maximum likelihood method is implemented to compute the MLE for the five models. By comparing AIC and BIC of the estimated regression models in Table 3.1, we can compare models and interpret the regression coefficients to examine how CAIX relates to CD8+ T cells.

Cross-K Function

On the cellular level, we categorize cells according to three phenotypes: CAIX cells, CD8+ T cells, and cells of other types. Taking the cell locations as a point pattern, we can treat the cellular level data as a multivariate marked point pattern whose marks are phenotype CAIX, CD8+, or others.

A polygonal boundary (the convex hull) is employed to define the window of the point pattern. The boundary is the farthest point in all directions from the center. Hence, the window (spatial domain) will change under different tissue samples. Let $K_{cd8,ca9}$ denote the cross-K function of CD8+ T cells in regard to CAIX cells. The theoretical cross-K function for a representative tissue sample $K_{cd8,ca9}$ is given by:

$$K_{cd8,ca9}(d) = \frac{E(\text{CAIX cell counts within a } d \text{ radius of circle of a CD8+ T cell})}{\lambda_{ca9}} \tag{3.1}$$

where λ_{ca9} is the intensity of CAIX over the window and $E(.)$ is the expectation function. Under CSR, $K_{cd8,ca9} = K_{ca9,cd8} = \pi d^2$.

From formula (2.26), on a representative tissue sample with M CAIX cell counts and N CD8+ T cell counts over the spatial domain A, the estimated Ripley's cross-K function for an individual tissue sample is defined as:

$$\hat{K}_{cd8,ca9}(d) = \frac{1}{\hat{\lambda}_{ca9}\hat{\lambda}_{cd8}A} \sum_m \sum_n \omega_{cd8_n,ca9_m} \, I(d_{cd8_n,ca9_m} < d) \tag{3.2}$$

where $d_{cd8_n,ca9_m}$ is the distance between nth $(n = 1, ...N)$ location of CD8+ T cells and mth $(m = 1, ...M)$ location of CAIX cells, and $\omega_{cd8_n,ca9_m}$ is the edge correction function. Since cores on TMAs are representative tissues taken from donor blocks, Ripley's isotropic correction estimate is appropriate to correct edge effects, which is

defined by the proportion of the circumference of a circle of nth location of CD8+ T cell within a search circle that lies inside the window. The indicator function $I(\,) = 1$ if $d_{cd8_n,ca9_m} < d$ is true, otherwise zero.

We then compare $\hat{K}_{cd8,ca9}(d)$ with the theoretical cross-K functions $K_{cd8,ca9}(d) = \pi d^2$ under CSR to test the independence between CAIX cells and CD8+ T cells. A clustered point pattern at a distance d is indicated by $\hat{K}_{cd8,ca9}(d) > K_{cd8,ca9}(d)$ and a regulated and dispersed point pattern at a distance d is suggested by $\hat{K}_{cd8,ca9}(d) < K_{cd8,ca9}(d)$. If $\hat{K}_{cd8,ca9}(d) \approx K_{cd8,ca9}(d)$, CAIX and CD8+ T cells are independent processes.

Aggregate Simulation Envelopes for Cross-K Function

As we discussed in Chapter 2, Monte Carlo simulation envelopes can be employed to assess the cross-K functions [23]. Suppose that there are T independent simulations under CSR with estimated cross-K functions $\hat{K}^{(t)}(d)$ for tth ($t = 1, 2, ..., T$) replication. For a representative tissue sample, the upper and lower envelope bounds are:

$$\begin{aligned} u(d) &= \max_{t\in\{1,...,T\}} \hat{K}^{(t)}(d) \\ l(d) &= \min_{t\in\{1,...,T\}} \hat{K}^{(t)}(d) \end{aligned} \tag{3.3}$$

The intervals between $u(d)$ and $l(d)$ for a given distance d constitute the simulation envelope for this representative tissue sample.

Our study consists of S tissue samples and therefore there is an envelope for each individual tissue sample image. Let $u^{(s)}(d)$ and $l^{(s)}(d)$ denote upper and lower envelope bounds for the sth ($s = 1, 2, .., S$) tissue sample. In their entirety, the simulated upper

and lower bounds for the aggregate envelope are:

$$\begin{aligned} U(d) &= \max_{s\in\{1,\ldots,S\}} u^{(s)}(d) \\ L(d) &= \min_{s\in\{1,\ldots,S\}} l^{(s)}(d) \end{aligned} \tag{3.4}$$

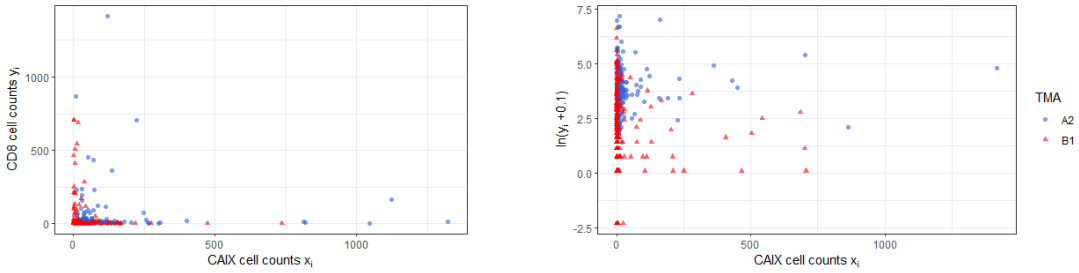
The intuition of aggregated envelope bound is straightforward. For any given distance d, the maximum value of all the simulated upper bounds turns into the overall upper bound; whereas, the minimum value of all the simulated lower bounds acts as the overall lower bound for the aggregate envelope. The aggregate envelope bounds for all the tissue samples are now $[L(d), U(d)]$.

We can conclude about the point pattern by examining whether the estimated cross-K function deviates from the envelopes $[L(d), U(d)]$. If the estimated cross-K functions are within $[L(d), U(d)]$ for a given distance d, CAIX cells and CD8+ T cells are independent. On the contrary, if the estimated cross-K functions lie outside $[L(d), U(d)]$ for a given distance d, there exists an dependence between CAIX cells and CD8+ T cells.

3.3 Discussion

The stroma, which is a component of TME to provide nutrients to cell proliferation, has a distinct structure from the tumor. We therefore, will further distinguish the results on tumor and stroma, respectively.

Figure 3.2 provides the scatter plots of CAIX and CD8+ T cell counts. It is difficult to observe a relation between CAIX and CD8+ cell counts on Figure 3.2(a). After taking the logarithm of CD8+ T cell count plus 0.1 to allow for zero cell count, neither positive or negative relationship is observed on Figure 3.2(b). We start with the



(a) Scatter plot of x_i and y_i (b) Scatter plot of x_i and $ln(y_i + 0.1)$

Figure 3.2: Scatter plot of CAIX cell counts x_i and CD8 cell counts y_i, and scatter plot CAIX cell counts x_i and natural log of CD8 cell counts y_i plus 0.1.

computation of Pearson correlation coefficients between CAIX cell counts and CD8+ T cell counts. The computed values with 95% confidence intervals are presented in Table 3.2. We can notice a variation between TMAs. The 95% confidence intervals on TMA A2 all contain 0; whereas the 95% confidence intervals on TMA B1, TMA B1 tumors, and TMA B1 stroma are all located underneath 0. There exists a negative correlation between CAIX and CD8+ T cells on TMA B1 and there appears to be a TMA effect.

The Spearman correlation coefficients are also computed and presented in Table 3.2, whose 95% confidence intervals cover 0 in all tissue samples, except TMA B1 stroma. There is weak evidence for a monotonic relationship between CAIX and CD8+ T cell counts.

The Poisson, Negative Binomial, and Zero-inflated Poisson Model Selection

We apply five count approach methods to model the CD8+ T cell counts per tissue core. As described in Table 3.1, these models are the Poisson, negative binomial(NB) and zero-inflated Poisson (ZIP). The results of Pearson correlation coefficients suggest a potential TMA fixed effect. A variety of models are obtained when taking TMA fixed effects into account. Based on the tissue category, the alternative samples are entire

Table 3.2: 95% Bootstrap confidence intervals for correlation between CAIX and CD8+ T cell counts for different tissue samples

Pearson correlation				
TMA	A2		B1	
	0.0282		-0.0788	
	(-0.0737, 0.1212)		(-0.1076, -0.0408)	
Tissue	Tumors (n=177)	Stroma (n=177)	Tumors (n=208)	Stroma (n=208)
	0.077	0.0374	-0.0437	-0.0734
	(-0.1032, 0.2313)	(-0.0651, 0.1299)	(-0.0746, -0.0094)	(-0.1075, -0.0203)

Spearman correlation				
TMA	A2		B1	
	0.05567		-0.1533	
	(-0.0918, 0.2058)		(-0.2886, -0.0184)	
Tissue	Tumors (n=177)	Stroma (n=177)	Tumors (n=208)	Stroma (n=208)
	0.0863	0.0773	-0.0194	-0.2477
	(-0.0654, 0.2408)	(-0.0728, 0.2292)	(-0.1598, 0.1220)	(-0.2850, -0.0119)

COEUR, tumor tissue, or stromal tissues. Applying the *glmmTMB* R package [12] with constant terms, the AIC, BIC, and -2loglikelihood are presented in Tables 3.3 to 3.5.

The Poisson model has the largest AIC and BIC. In addition, the overdispersion test is applied to test the null hypothesis of no dispersion against the alternative of overdispersion. The computed p-values of the overdispersion test are 9.065×10^{-4} (for overall cell counts), 4.052×10^{-3} (for cell counts in tumors), and 5.04×10^{-3} (for cell counts in stroma), leading to a strong rejection of the null. Thus, ZIP and NB models are employed to allow for over-dispersion. ZIP models substantially reduce AIC and BIC compared to the Poisson. Of the remaining models, the NB with TMA fixed effect model yields the smallest AIC, BIC and the lowest -2log-likelihood, implying that this is the best model for the cell counts of these models considered.

Regression coefficients, 95% confidence intervals, and p-values of NB models are available in Table 3.6. It is observed that TMA has a significant impact on CD8+ T

Table 3.3: Model comparison for all the COEUR tissue samples

Criteria	Poisson	ZIP	ZIP with FE	NB	NB with FE
AIC	47847.92	47013.47	40599.816	3794.294	3715.946
BIC	47855.83	47025.33	40619.582	3806.154	3735.712
-2log-likelihood	47843.92	47007.47	40589.816	3788.294	3705.946
Df of resid	383	382	380	382	380

Table 3.4: Model comparison for tumors

Criteria	Poisson	ZIP	ZIP with FE	NB	NB with FE
AIC	16984.34	15458.909	15361.89	2844.609	2842.507
BIC	16992.25	15470.768	15381.656	2856.468	2862.273
-2log-likelihood	16980.34	15452.909	15351.89	2838.609	2832.5
Df of resid	383	382	380	382	380

cell count. The estimates of TMA coefficients depend on the tissue category. The interaction between CAIX and TMA B1 is significant at 1% significance level on overall tissue and stroma, and 10% significance level on tumors. Based on Table 3.1, the explicit model specifications are $(ln(y_i|A2) = \beta_0 + \beta_1 x_i)$ at TMA A2 and $(ln(y_i|B1) = \beta_0 + \beta_2 + (\beta_1 + \beta_3)x_i)$ at TMA B1. The confidence interval of the CAIX coefficient is conditional on the TMA. Taking the interaction between CAIX and TMA into account, the 95% confidence interval of CAIX is provided in Table 3.7. It is observed that the confidence interval for tissue samples on TMA B1 is smaller than 0, suggesting a negative relationship between CAIX and CD8+ T cells; whereas the confidence intervals of hypoxia on TMA A2 contain 0, meaning that hypoxia is likely to be insignificant on TMA A2.

The NB model is preferred on the condition that overdispersion is caused by the unobserved heterogeneity. We explore the dependence between CAIX and CD8+ T cells by taking spatial information into consideration in the next section. Overall, NB models show that TMAs have a significant CD8+ T cell count.

Table 3.5: Model comparison for stroma

Criteria	Poisson	ZIP	ZIP with FE	NB	NB with FE
AIC	38540.22	37401.19	29926.946	3464.317	3322.619
BIC	38548.13	37413.05	29946.712	3476.176	3342.385
-2log-likelihood	38536.22	37395.19	29916.946	3458.317	3312.619
Df of resid	383	382	380	382	380

Table 3.6: Estimated regression coefficients of NB model for different tissue samples

Covariate	Coefficients	Estimate	95% confidence intervals	p-value
Overall (over-dispersion parameter $\alpha = 1.5205$)				
Intercept	β_0	4.4813	(4.2900, 4.6726)	0.0000
CAIX	β_1	0.0004	(-0.0009, 0.0004)	0.5538
TMA B1	β_2	-1.0699	(-1.3296, -0.8102)	0.0000
CAIX×TMA B1	β_3	-0.0032	(-0.0051, -0.0013)	0.0010
Tumor (over-dispersion parameter $\alpha = 2.2172$)				
Intercept	β_0	2.9919	(2.7580, 3.2258)	0.0000
CAIX	β_1	0.0089	(-0.0095, 0.0273)	0.3448
TMA B1	β_2	-0.2500	(-0.5634, 0.0634)	01179
CAIX × TMA B1	β_3	-0.0289	(-0.0611, 0.0033)	0.0780
Stroma (over-dispersion parameter $\alpha = 1.4882$)				
Intercept	β_0	4.2097	(4.0204, 4.3991)	0.0000
CAIX	β_1	0.0006	(-0.0008, 0.0019)	0.394
TMA B1	β_2	-1.5474	(-1.8063, -1.2885)	0.0000
CAIX×TMA B1	β_3	-0.0040	(-0.0062, -0.0018)	0.0003

Table 3.7: 95% Confidence intervals of hypoxia conditional on TMA from the NB model regressions

Samples	TMA A2	TMA B1
Overall	(-0.00087, 0.00162)	(-0.00425, -0.00137)
Tumor	(-0.00955, 0.02732)	(-0.04644, 0.00634)
Stroma	(-0.00076, 0.00193)	(-0.00511, -0.00170)

Plots of Cross-K Functions

The estimated cross-K functions of CAIX and CD8+ T cells are computed. To provide variability, Monte Carlo simulation is used to derive the combined simulated envelopes under CSR. We run 50 to 999 Monte Carlo simulations, observe the simulated envelopes, and label the core that has a visual stable envelope. We find that a minimum intensity 4×10^{-5} is set to keep the stability of Monte Carlo simulation under the law of large numbers. We compute the cross-K function for the cores with intensity greater than 4×10^{-5}. When 99 Monte Carlo simulations are employed, the probability that the estimated cross-K function lying outside the envelope is 2%. Hence, the 98% envelopes are simulated and aggregated so that they align with the plots for estimated cross-K functions of CAIX and CD8+ T cells in Figure 3.3.

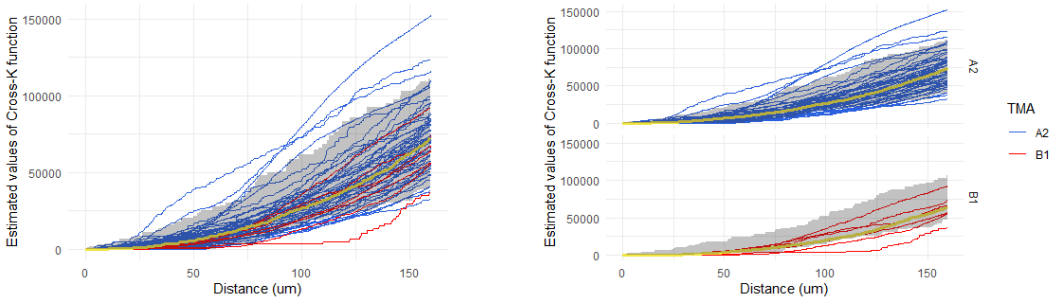
As we can observe, there are more tissue cores on TMA A2 after putting a threshold for intensity. When all the sample tissues are taken in account, the estimated cross-K functions $\hat{K}_{cd8,ca9}(d)$ on Figure 3.3(a) show that most tissue cores lie within the envelopes. We see 4 tissue cores from TMA A2 are observed above the envelopes, providing evidence in favor of a clustering pattern. The only estimated cross-K function located under the envelopes is a tissue core from TMA B1, which CAIX and CD8+ T cells are regulated in their pattern.

Since tissue category is an important feature for TME, we estimate cross-K functions of CAIX and CD8+ T cells by tissue category, respectively. Figure 3.4 illustrates the estimated cross-K functions of CAIX and CD8+ T cells by tissue category.

Figure 3.4(a) and Figure 3.4(b) focus on tumor regions. It is found that the estimated cross-K functions of 3 cores on TMA A2 are located in the envelopes and there is no dependence between CAIX and CD8+ T cells. On the other hand, it is observed

that 2 cores on TMA A2 and 1 core on TMA B1 deviate from the envelopes, which means CAIX and CD8+ T cells are regulated on these 3 cores.

With respect to stroma plotted on Figure 3.4(c) and Figure 3.4(d), the estimated cross-K functions are observed spanning within and outside the envelopes. The dependence between CAIX and CD8+ T cells on TMA A2 is ambiguous. The evidence is inconclusive. Only 2 cores on TMA B1 are used to estimate the cross-K functions with no evidence of dependence between CAIX and CD8+ T cells.

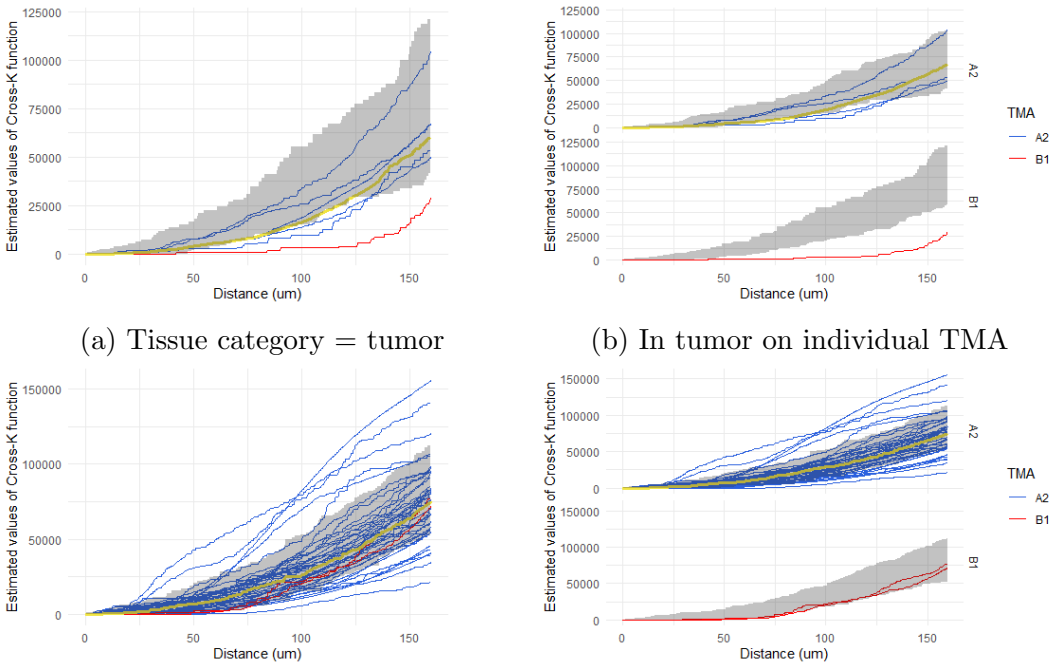


(a) All the tissues (b) All tissues by individual TMA

Figure 3.3: Estimated cross-K functions of CAIX and CD8+ T cells with aggregate envelopes (gray area). The yellow curve is the median value of all the envelopes for overall cores.

3.4 Concluding Remarks

The evidence indicates a strong TMA effect. The relation between CAIX and CD8+ T cells highly depends on TMA. On TMA B1, the 95% confidence intervals of correlation suggest CAIX and CD8+ T cells are negatively corrected. The negative relation is observed on the overall (and stroma) of TMA B1 in the negative binomial model. The cross-K functions support this negative correlation on TMA B1. The cross-K function plots discover that CAIX and CD8+ T cells are regulated, that is to say, repel each other at TMA B1. Tissue category has an influential impact on cross-K functions.



(a) Tissue category = tumor

(b) In tumor on individual TMA

(c) Tissue category = stroma

(d) In stroma on individual TMA

Figure 3.4: Estimated cross-K functions of CAIX and CD8+ T cells with envelopes across tissue samples and TMAs. The yellow curve is the median value of all the envelopes presented in the figure.

CAIX and CD8+ T cells on tumors are more regulated, compared with TMA B1 stroma.

But on TMA A2, the relation between CAIX and CD8+ T cells are inconclusive. The range of 95% confidence intervals of correlation contains 0 and the estimated cross-K functions spanning inside and outside the Monte Carol simulated envelopes. Thus the level of noise produces any statement as the signal. Clustering, regulation, and independence are all observed on TMA A2. Tissue category also appears to impact the point the spatial structure of the pattern. Clustering is not observed on TMA A2 tumors. CAIX and CD8+ T cells on TMA A2 are likely to be regulated on tumor tissue.

To summarize, a pronounced heterogeneity is found in the samples. TMA and tissue category have a strong impact on the relation between CAIX and CD8+ T cells.

We see a variability between TMA A2 and B1. The relationship between CAIX and CD8+ T cells is inconclusive on TMA A2. But a negative relationship is observed at TMA B1. Tissue category is an important factor. CAIX is negatively associated with CD8+ T cells on TMA B1 stroma, which is not observed on TMA B1 tumor in the negative binomial model. When taking the spatial domain into account, we observe that CAIX and CD8+ T cells within the tissue core repel each other in the tumors of TMA B1.

Chapter 4

Future Work and Conclusions

In this paper, we investigate the relationship between CAIX and CD8+ T cells. There may be some evidence that CAIX and CD8+ T cells are correlated, depending on TMAs and tissue category. This evidence is not strong. The negative correlation between CAIX cell counts is observed on TMA B1. The negative binomial model indicates a negative relationship between CAIX and CD8+ T cell counts on the TMA B1 stroma. Additionally, the estimated cross-K function on TMA B1 tumor demonstrates that CAIX repels CD8+ T cells within TME.

Further work can be conducted to improve both the dataset and statistical analysis. NB models hold a superior performance when handling overdispersion induced by unobserved heterogeneity for these data. The variability between TMAs can be attributed to the heterogeneity in tissue cores. One aspect of future work is to incorporate more characteristics of the tissue in the NB model, such as donor age [13], tumor grade and genetics [62] to reduce the heterogeneity.

Survival rates of ovarian cancer are major interest and the other potential extension for future analysis is to include survival data. The survival model is a potential approach to analyze cancer survival in relation to CAIX and CD8+ T cells. If we track

patients and keep a record of a donor's tissue change over multiple years, it is useful to probe how the dependence of CAIX and CD8+ T cells varies on a time frame and how it relates to survival rates.

The other possible improvement of the thesis is to account for the measurement errors in cell counts by developing new joint negative binomial and survival models. A Bayesian hierarchical model can be built to join these models.

When multiple-year tissue data is available, the other possible approach is spatial-temporal point pattern analysis by considering longitudinal analysis of point patterns. Such an analysis can give us more insight about whether the clustering, independence, and regulated point patterns change over years.

Bibliography

[1] About ovarian cancer. http://ovariancanada.org/About-Ovarian-Cancer. Accessed June 12, 2021.

[2] Cytotoxic t cell overview. https://www.thermofisher.com/ca/en/home/life-science/cell-analysis/cell-analysis-learning-center/immunology-at-work/cytotoxic-t-cell-overview.html. Accessed July 12, 2021.

[3] Ovary - epithelial carcinoma. http://http://www.bccancer.bc.ca/health-professionals/clinical-resources/cancer-management-guidelines/gynecology/ovary-epithelial-carcinoma. Accessed August 14, 2021.

[4] Survival statistics for ovarian cancer. https://www.cancer.ca/en/cancer-information/cancer-type/ovarian/prognosis-and-survival/survival-statistics/?region=on. Access July 1, 2021.

[5] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer, 1998.

[6] M. Bache, M. Kappler, H. Said, A. Staab, and D. Vordermark. Detection and specific targeting of hypoxic regions within solid tumors: current preclinical and clinical strategies. *Current medicinal chemistry*, 15(4):322–338, 2008.

[7] A. Baddeley, E. Rubak, and R. Turner. *Spatial point patterns: methodology and applications with R*. CRC press, 2015.

[8] P. Bankhead, M. B. Loughrey, J. A. Fernández, Y. Dombrowski, D. G. McArt, P. D. Dunne, S. McQuaid, R. T. Gray, L. J. Murray, H. G. Coleman, et al. Qupath: Open source software for digital pathology image analysis. *Scientific reports*, 7(1):1–7, 2017.

[9] N. Bannoud, T. Dalotto-Moreno, L. Kindgard, P. A. García, A. G. Blidner, K. V. Mariño, G. A. Rabinovich, and D. O. Croci. Hypoxia supports differentiation of terminally exhausted cd8 t cells. *Frontiers in Immunology*, 12, 2021.

[10] I. B. Barsoum, C. A. Smallwood, D. R. Siemens, and C. H. Graham. A mechanism of hypoxia-mediated escape from adaptive immunity in cancer cells. *Cancer research*, 74(3):665–674, 2014.

[11] S. Barua, P. Fang, A. Sharma, J. Fujimoto, I. Wistuba, A. U. Rao, and S. H. Lin. Spatial interaction of tumor cells and regulatory t cells correlates with survival in non-small cell lung cancer. *Lung Cancer*, 117:73–79, 2018.

[12] M. E. Brooks, K. Kristensen, K. J. Van Benthem, A. Magnusson, C. W. Berg, A. Nielsen, H. J. Skaug, M. Machler, and B. M. Bolker. glmmtmb balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R journal*, 9(2):378–400, 2017.

[13] Canadian Cancer Statistics Advisory Committee. Canadian cancer statistics 2019. `http://cancer.ca/Canadian-Cancer-Statistics-2019-EN`, September 2019. Access June 20, 2021.

[14] Cancer Research UK. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/ovarian-cancer/incidence#ref-8. Accessed June 20, 2021.

[15] J. L. Carrasco. A generalized concordance correlation coefficient based on the variance components generalized linear mixed models for overdispersed count data. *Biometrics*, 66(3):897–904, 2010.

[16] M. Castells, B. Thibault, J.-P. Delord, and B. Couderc. Implication of tumor microenvironment in chemoresistance: tumor-associated stromal cells protect tumor cells from cell death. *International journal of molecular sciences*, 13(8):9545–9571, 2012.

[17] S. K. Chia, C. C. Wykoff, P. H. Watson, C. Han, R. D. Leek, J. Pastorek, K. C. Gatter, P. Ratcliffe, and A. L. Harris. Prognostic significance of a novel hypoxia-regulated marker, carbonic anhydrase ix, in invasive breast carcinoma. *Journal of Clinical Oncology*, 19(16):3660–3668, 2001.

[18] M. Choschzick, E. Oosterwijk, V. Müller, L. Woelber, R. Simon, H. Moch, and P. Tennstedt. Overexpression of carbonic anhydrase ix (caix) is an independent unfavorable prognostic marker in endometrioid ovarian cancer. *Virchows Archiv*, 459(2):193–200, 2011.

[19] N. Colombo, T. Van Gorp, G. Parma, F. Amant, G. Gatta, C. Sessa, and I. Vergote. Ovarian cancer. *Critical reviews in oncology/hematology*, 60(2):159–179, 2006.

[20] E. Committee on the State of the Science in Ovarian Cancer Research; Board on Health Care Services; Institute of Medicine; National Academies of Sciences and

Medicine. *Ovarian Cancers: Evolving Paradigms in Research and Care*. National Academies Press, 2016.

[21] F. Dayan, N. M. Mazure, M. C. Brahimi-Horn, and J. Pouysségur. A dialogue between the hypoxia-inducible factor and the tumor microenvironment. *Cancer Microenvironment*, 1(1):53–68, 2008.

[22] P. J. Diggle. Spatio-temporal point processes: methods and applications. *Monographs on Statistics and Applied Probability*, 107:1, 2006.

[23] P. J. Diggle. *Statistical analysis of spatial and spatio-temporal point patterns*. CRC press, 2013.

[24] P. M. Dixon. Ripley's k function. *Encyclopedia of environmetrics*, 3:1796, 2001.

[25] M. Duechler, L. Peczek, M. Szubert, and J. Suzin. Influence of hypoxia inducible factors on the immune microenvironment in ovarian cancer. *Anticancer research*, 34(6):2811–2819, 2014.

[26] A. C. Gatrell, T. C. Bailey, P. J. Diggle, and B. S. Rowlingson. Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British geographers*, pages 256–274, 1996.

[27] R. J. Gillies, N. Raghunand, G. S. Karczmar, and Z. M. Bhujwalla. Mri of the tumor microenvironment. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 16(4):430–450, 2002.

[28] Y. Gropper, T. Feferman, T. Shalit, T.-M. Salame, Z. Porat, and G. Shakhar. Culturing ctls under hypoxic conditions enhances their cytolysis and improves their anti-tumor function. *Cell reports*, 20(11):2547–2555, 2017.

[29] J. Hinde and C. G. Demetrio. Overdispersion: models and estimation. *Computational statistics and data analysis*, 27(2):151–170, 1998.

[30] I. J. Hoogsteen, H. A. Marres, K. I. Wijffels, P. F. Rijken, J. P. Peters, F. J. van den Hoogen, E. Oosterwijk, A. J. van der Kogel, and J. H. Kaanders. Colocalization of carbonic anhydrase 9 expression and cell proliferation in human head and neck squamous cell carcinoma. *Clinical cancer research*, 11(1):97–106, 2005.

[31] A. Horiuchi, T. Hayashi, N. Kikuchi, A. Hayashi, C. Fuseya, T. Shiozawa, and I. Konishi. Hypoxia upregulates ovarian cancer invasiveness via the binding of hif-1α to a hypoxia-induced, methylation-free hypoxia response element of s100a4 gene. *International journal of cancer*, 131(8):1755–1767, 2012.

[32] M. R. Horsman and P. Vaupel. Pathophysiological basis for the formation of the tumor microenvironment. *Frontiers in oncology*, 6:66, 2016.

[33] M.-C. Hu, M. Pavlicova, and E. V. Nunes. Zero-inflated and hurdle models of count data with extra zeros: examples from an hiv-risk reduction intervention trial. *The American journal of drug and alcohol abuse*, 37(5):367–375, 2011.

[34] C. A. Janeway Jr, P. Travers, M. Walport, and M. J. Shlomchik. *Immunobiology The Immune System in Health and Disease*. Garland Science, New York, 5 edition, 2001.

[35] Q. Ke and M. Costa. Hypoxia-inducible factor-1 (hif-1). *Molecular pharmacology*, 70(5):1469–1480, 2006.

[36] H.-J. Kim and H. Cantor. Cd4 t-cell subsets and tumor immunity: the helpful and the not-so-helpful. *Cancer immunology research*, 2(2):91–98, 2014.

[37] K.-S. Kim, S. Sengupta, M. Berk, Y.-G. Kwak, P. F. Escobar, J. Belinson, S. C. Mok, and Y. Xu. Hypoxia enhances lysophosphatidic acid responsiveness in ovarian cancer cells and lysophosphatidic acid induces ovarian tumor metastasis in vivo. *Cancer research*, 66(16):7983–7990, 2006.

[38] D. Lambert. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992.

[39] R. K. Lodwick, C. A. Sabin, K. Porter, B. Ledergerber, A. Van Sighem, A. Cozzi-Lepri, P. Khaykin, A. Mocroft, L. Jacobson, S. De Wit, et al. Death rates in hiv-positive antiretroviral-naive patients with cd4 count greater than 350 cells per microl in europe and north america: a pooled cohort observational study. *Lancet (London, England)*, 376(9738):340–345, 2010.

[40] J. A. Loncaster, A. L. Harris, S. E. Davidson, J. P. Logue, R. D. Hunter, C. C. Wycoff, J. Pastorek, P. J. Ratcliffe, I. J. Stratford, and C. M. West. Carbonic anhydrase (ca ix) expression, a potential new intrinsic marker of hypoxia: correlations with tumor oxygen measurements and prognosis in locally advanced carcinoma of the cervix. *Cancer research*, 61(17):6394–6399, 2001.

[41] N. R. Maimela, S. Liu, and Y. Zhang. Fates of cd8+ t cells in tumor microenvironment. *Computational and structural biotechnology journal*, 17:1–13, 2019.

[42] Y. Masugi, T. Abe, A. Ueno, Y. Fujii-Nishimura, H. Ojima, Y. Endo, Y. Fujita, M. Kitago, M. Shinoda, Y. Kitagawa, et al. Characterization of spatial distribution of tumor-infiltrating cd8+ t cells refines their prognostic utility for pancreatic cancer survival. *Modern Pathology*, 32(10):1495–1507, 2019.

[43] R. Matkowski, I. Gisterek, A. Halon, A. Lacko, K. Szewczyk, U. Staszek, M. Pudelko, B. Szynglarewicz, J. Szelachowska, A. Zolnierek, et al. The prognostic role of tumor-infiltrating cd4 and cd8 t lymphocytes in breast cancer. *Anticancer research*, 29(7):2445–2451, 2009.

[44] A. Maydeu-Olivares and C. Garcia-Forero. Goodness-of-fit testing. *International encyclopedia of education*, 7(1):190–196, 2010.

[45] P. McCullagh and J. Nelder. Generalized linear models ii, 1989.

[46] P. C. McDonald, J.-Y. Winum, C. T. Supuran, and S. Dedhar. Recent developments in targeting carbonic anhydrase ix for cancer therapeutics. *Oncotarget*, 3(1):84, 2012.

[47] E. N. McNamee, D. K. Johnson, D. Homann, and E. T. Clambey. Hypoxia and hypoxia-inducible factors as regulators of t cell development, differentiation, and function. *Immunologic research*, 55(1):58–70, 2013.

[48] R. H. Myers, D. C. Montgomery, G. G. Vining, and T. J. Robinson. *Generalized linear models: with applications in engineering and the sciences*, volume 791. John Wiley & Sons, 2012.

[49] Y. Naito, K. Saito, K. Shiiba, A. Ohuchi, K. Saigenji, H. Nagura, and H. Ohtani. Cd8+ t cells infiltrated within cancer cell nests as a prognostic factor in human colorectal cancer. *Cancer research*, 58(16):3491–3494, 1998.

[50] H. Nakamura, H. Saji, A. Ogata, M. Hosaka, M. Hagiwara, N. Kawasaki, C. Konaka, and H. Kato. Immunologic parameters as significant prognostic factors in lung cancer. *Lung Cancer*, 37(2):161–169, 2002.

[51] M. Z. Noman, M. Hasmim, Y. Messai, S. Terry, C. Kieda, B. Janji, and S. Chouaib. Hypoxia: a key player in antitumor immune response. a review in the theme: cellular responses to hypoxia. *American Journal of Physiology-Cell Physiology*, 309(9):C569–C579, 2015.

[52] D. W. Osgood. Poisson-based regression analysis of aggregate crime rates. *Journal of quantitative criminology*, 16(1):21–43, 2000.

[53] B. Pittman, E. Buta, S. Krishnan-Sarin, S. S. O'Malley, T. Liss, and R. Gueorguieva. Models for analyzing zero-inflated and overdispersed count data: an application to cigarette and marijuana use. *Nicotine and Tobacco Research*, 22(8):1390–1398, 2020.

[54] P. Prasad, C. R. Gordijo, A. Z. Abbasi, A. Maeda, A. Ip, A. M. Rauth, R. S. DaCosta, and X. Y. Wu. Multifunctional albumin–mno2 nanoparticles modulate solid tumor microenvironment by attenuating hypoxia, acidosis, vascular endothelial growth factor and enhance radiation response. *ACS nano*, 8(4):3202–3212, 2014.

[55] J. Qin, Y. Liu, Y. Lu, M. Liu, M. Li, J. Li, and L. Wu. Hypoxia-inducible factor 1 alpha promotes cancer stem cells-like properties in human ovarian cancer cells by upregulating sirt1 expression. *Scientific reports*, 7(1):1–12, 2017.

[56] G. Rodrıguez. Models for count data with overdispersion. *Addendum to the WWS*, 509, 2013.

[57] C. E. Rose, S. W. Martin, K. A. Wannemuehler, and B. D. Plikaytis. On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *Journal of biopharmaceutical statistics*, 16(4):463–481, 2006.

[58] R. Simon, M. Mirlacher, and G. Sauter. Tissue microarrays. *Biotechniques*, 36(1):98–105, 2004.

[59] D. E. Swinson, J. L. Jones, D. Richardson, C. Wykoff, H. Turley, J. Pastorek, N. Taub, A. L. Harris, and K. J. O Byrne. Carbonic anhydrase ix expression, a novel surrogate marker of tumor hypoxia, is associated with a poor prognosis in non-small-cell lung cancer. *Journal of Clinical Oncology*, 21(3):473–482, 2003.

[60] M. Thiel, C. C. Caldwell, S. Kreth, S. Kuboki, P. Chen, P. Smith, A. Ohta, A. B. Lentsch, D. Lukashev, and M. V. Sitkovsky. Targeted deletion of hif-1α gene in t cells prevents their inhibition in hypoxic inflamed tissues and improves septic mice survival. *PLoS One*, 2(9):e853, 2007.

[61] R. Vuillefroy de Silly, P.-Y. Dietrich, and P. R. Walker. Hypoxia and antitumor cd8+ t cells: An incompatible alliance? *Oncoimmunology*, 5(12):e1232236, 2016.

[62] R. Vuillefroy de Silly, L. Ducimetière, C. Yacoub Maroun, P.-Y. Dietrich, M. Derouazi, and P. R. Walker. Phenotypic switch of cd8+ t cells reactivated under hypoxia toward il-10 secreting, poorly proliferative effector cells. *European journal of immunology*, 45(8):2263–2275, 2015.

[63] C. WalshJoseph, C. KolbHartmuth, et al. The clinical importance of assessing tumor hypoxia: relationship of tumor hypoxia to prognosis and therapeutic opportunities. *Antioxidants & redox signaling*, 2014.

[64] T. Whiteside. The tumor microenvironment and its role in promoting tumor growth. *Oncogene*, 27(45):5904–5912, 2008.

[65] C. M. Wilson, O. E. Ospina, M. K. Townsend, J. Nguyen, C. Moran Segura, J. M. Schildkraut, S. S. Tworoger, L. C. Peres, and B. L. Fridley. Challenges and

opportunities in the statistical analysis of multiplex immunofluorescence data. *Cancers*, 13(12):3031, 2021.

[66] C. Wong, T. L. Wellman, and K. M. Lounsbury. Vegf and hif-1α expression are increased in advanced stages of epithelial ovarian cancer. *Gynecologic oncology*, 91(3):513–517, 2003.

[67] F. Yu, S. B. White, Q. Zhao, and F. S. Lee. Hif-1α binding to vhl is regulated by stimulus-sensitive proline hydroxylation. *Proceedings of the National Academy of Sciences*, 98(17):9630–9635, 2001.

[68] K. Zhang, X. Kong, G. Feng, W. Xiang, L. Chen, F. Yang, C. Cao, Y. Ding, H. Chen, M. Chu, et al. Investigation of hypoxia networks in ovarian cancer via bioinformatics analysis. *Journal of ovarian research*, 11(1):1–11, 2018.

[69] L. Zhang, J. R. Conejo-Garcia, D. Katsaros, P. A. Gimotty, M. Massobrio, G. Regnani, A. Makrigiannakis, H. Gray, K. Schlienger, M. N. Liebman, et al. Intratumoral t cells, recurrence, and survival in epithelial ovarian cancer. *New England journal of medicine*, 348(3):203–213, 2003.

[70] Y. Zhang, M. Coleman, and R. A. Brekken. Perspectives on hypoxia signaling in tumor stroma. *Cancers*, 13(12):3070, 2021.

[71] X. Zheng, A. Weigert, S. Reu, S. Guenther, S. Mansouri, B. Bassaly, S. Gattenlöhner, F. Grimminger, S. S. Pullamsetti, W. Seeger, et al. Spatial density and distribution of tumor-associated macrophages predict survival in non–small cell lung carcinoma. *Cancer Research*, 80(20):4414–4425, 2020.