

# Spatial analysis of county level Colorectal Cancer incidence in Texas and Bayesian analysis of association with county level smoking and education demographic data

Khush Patel, MD<sup>1\*</sup>

<sup>1</sup>Division of Epidemiology, Human Genetics and Environmental Science, University of Texas School of Public Health, 1200 Herman Pressler, Houston, TX 77030, USA

Email: Khush.A.Patel@uth.tmc.edu

\*corresponding author

## Abstract

**Background:** To conduct spatial analysis of county level Colorectal Cancer Incidence in Texas and perform Bayesian analysis using integrated nested Laplace approximation of county level smoking demographic.

**Methods:** County level colorectal cancer incidence data for Texas were obtained for the years 2009 to 2018 were obtained from the Texas Cancer Registry, Cancer Epidemiology and Surveillance Branch, Texas Department of State Health Services. Corresponding years demographic data at county level was obtained for smoking, bachelor's degree as a proxy to colorectal cancer screening awareness and race. Exploratory analysis using Moran's I at global and local level was performed. Bayesian Poisson Model, Bayesian non-spatial model (iid model with non-spatial random effects) and Bayesian Spatial model (Conditional autoregressive (ICAR) model with spatial and non-spatial random effects) were fitted. We also developed a unique, novel statistical imputation method for suppressed cancer data not used before in any other spatial studies.

**Results:** We found local Moran's I to have positive spatial autocorrelation with value of +0.16 with p value <0.01. We found Conditional autoregressive (ICAR) model with spatial and non-spatial random effects fitted in INLA to be the model with lowest WAIC. All the models showed county level smoking data to be significantly associated with colorectal cancer incidence.

**Conclusion:** Colorectal cancer incidence showed significant autocorrelation amongst the county. County smoking data was significantly associated with Colorectal cancer incidence.

## Background

Colorectal cancer (CRC) is the third most common cancer[1, 2] in both males and females in the United States of America. The American Cancer Society's estimates for the number of colorectal cancer cases in the United States for 2021 are 149,500 cases[1]. These cancers can also be called colon cancer or rectal cancer, depending on where they start. Colon cancer and rectal cancer are often grouped together because they have many features in common. CRC incidence and mortality rates vary markedly around the world. Potentially modifiable behavior which are risk factors include smoking, unhealthy diet and obesity[3]. Other important factor is CRC screening rate. CRC screening has been significantly linked to education level in past studies[4, 5].

There is considerable variation in the rates of colorectal cancer incidence at the county level in state of Texas. The reported highest and lowest age-adjusted incidence rates for colorectal cancer from 2009 to

2018 for Texas at the county level were 14563 in and 16 per 100,000, respectively. Cancer incidence less than 16 were suppressed[6]. Relatively few studies have been conducted to examine how county level smoking status, education level and race information affects the geographic distribution of colorectal cancer incidence, and no study has been conducted to investigate this at the county level in the state of Texas. Identification of geographic patterns of colorectal cancer incidence could provide impetus to conduct further investigations and target health resources for prevention and treatment in specific geographic areas.

## Methods

### Data Description:

Geographic, Population, and Incidence Data: The case definition for this study was colorectal cancer incidence due to malignant neoplasm of colon, rectum and rectosigmoid location as listed by the International Classification of Diseases, 10th Revision (ICD-10)\* in the overall population of Texas for the years 2009-2018. The CRC incidence data at the county level for the years 2009 to 2018 were obtained from the Texas Cancer Registry, Cancer Epidemiology and Surveillance Branch, Texas Department of State Health Services[6]. The Texas Cancer Registry classifies cancer incidence data according to the Surveillance, Epidemiology, and End Results (SEER) definition of cancer incidence rate as the number of new cancers of a specific site/type occurring in a specified population during a year, usually expressed as the number of cancers per 100,000 population at risk. The Texas Cancer Registry is a statewide, population-based registry that collects high-quality, population-based data reported from various sources, including hospitals, cancer treatment centers, ambulatory surgery centers, pathology laboratories, and physician's offices through active and passive surveillance. It currently meets standards set by the National Program of Central Cancer Registries, Centers for Disease Control and Prevention for high-quality data, and is Gold-Certified by the North American Association of Central Cancer Registries. Age-adjusting takes the 2000 US population distribution and applies it to other time periods under consideration.

\*ICD codes used

Cancer causes	ICD-9	ICD-10
Colon excluding Rectum	153, 159.0	C18, C26.0
Rectum and Rectosigmoid Junction	154.0-154.1	C19-C20
Anus, Anal Canal and Anorectum	154.2-154.3, 154.8	C21

Population at risk (10 years population) : Total		267695761
Min Population	12521	Sterling
Max Population	43954456	Harris

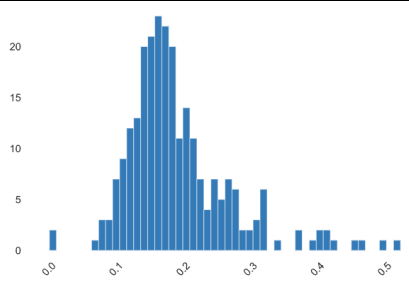
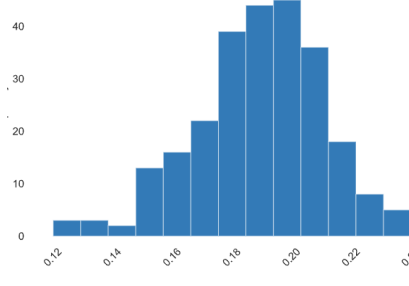
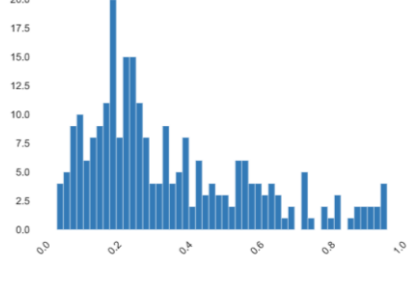
Total Age adjusted Cases in all races*		
Min Cases	16	Sherman
Max Cases	14563	Harris

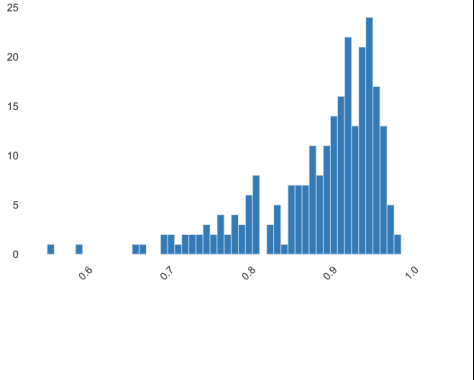
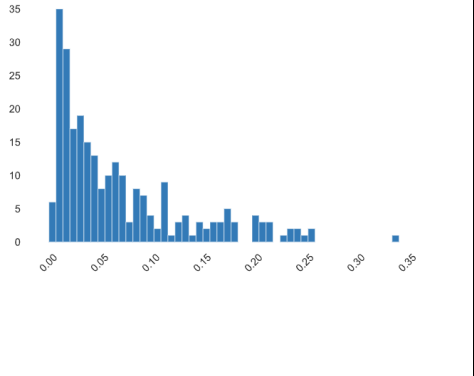
\*Rates/Counts are suppressed if less than 16 cases were reported in the specified category.

### Study variables:

We standardized all variables to have value between 0 and 1. We included county level smoking data. We included bachelor's degree as a proxy to colorectal cancer screening awareness as described above. African American, Caucasian, Hispanics were included as separate variables. The data was obtained from Texas Association of Counties[7].

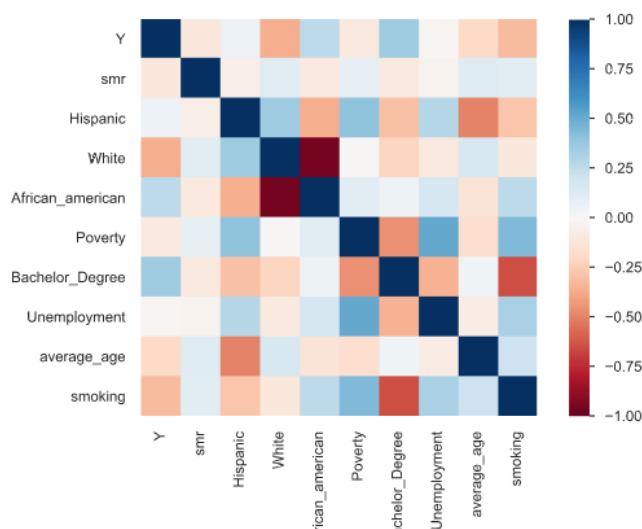
### Summary statistics for each county level feature to be analyzed.

Variable name	Summary Statistics		Histogram	
Percentage population having bachelor's degree:	Quantile statistics			
	Minimum	0		
	5-th percentile	0.09765		
	Q1	0.14225		
	median	0.173		
	Q3	0.218		
	95-th percentile	0.33135		
	Maximum	0.523		
	Range	0.523		
	Interquartile range (IQR)	0.07575		
Descriptive statistics				
Standard deviation		0.07801435786		
Coefficient of variation (CV)		0.4094057333		
Kurtosis		3.093371714		
Mean		0.1905551181		
Median Absolute Deviation (MAD)		0.036		
Skewness		1.430938611		
Sum		48.401		
Variance		0.006086240033		
Monotocity		Not monotonic		
Smoking percentage	Quantile statistics			
	Minimum	0.12		
	5-th percentile	0.15		
	Q1	0.18		
	median	0.19		
	Q3	0.21		
	95-th percentile	0.2235		
	Maximum	0.24		
	Range	0.12		
	Interquartile range (IQR)	0.03		
Descriptive statistics				
Standard deviation		0.02342203456		
Coefficient of variation (CV)		0.1233761256		
Kurtosis		0.2615485653		
Mean		0.1898425197		
Median Absolute Deviation (MAD)		0.01		
Skewness		-0.4292486321		
Sum		48.22		
Variance		0.0005485917027		
Monotocity		Not monotonic		
Hispanics	Quantile statistics			
	Minimum	0.0369		
	5-th percentile	0.081085		
	Q1	0.18585		
	median	0.2731		
	Q3	0.49165		
	95-th percentile	0.84268		
	Maximum	0.963		
	Range	0.9261		
	Interquartile range (IQR)	0.3058		
Descriptive statistics				
Standard deviation		0.2298000649		
Coefficient of variation (CV)		0.652815538		
Kurtosis		0.07552355779		
Mean		0.3520137795		
Median Absolute Deviation (MAD)		0.13125		
Skewness		0.950176235		
Sum		89.4115		
Variance		0.05280806981		
Monotocity		Not monotonic		

White	Quantile statistics		Descriptive statistics		
	Minimum	0.5524	Standard deviation	0.07505956616	
	5-th percentile	0.734475	Coefficient of variation (CV)	0.08454752991	
	Q1	0.857175	Kurtosis	2.230457464	
	median	0.9126	Mean	0.8877795276	
	Q3	0.941625	Median Absolute Deviation (MAD)	0.0367	
	95-th percentile	0.964805	Skewness	-1.459894509	
	Maximum	0.9865	Sum	225.496	
	Range	0.4341	Variance	0.005633938473	
	Interquartile range (IQR)	0.08445	Monotocity	Not monotonic	
African American	Quantile statistics		Descriptive statistics		
	Minimum	0	Standard deviation	0.06494715971	
	5-th percentile	0.008765	Coefficient of variation (CV)	0.9493668749	
	Q1	0.018375	Kurtosis	1.460020248	
	median	0.0438	Mean	0.06841102362	
	Q3	0.093575	Median Absolute Deviation (MAD)	0.02945	
	95-th percentile	0.211035	Skewness	1.399387791	
	Maximum	0.3414	Sum	17.3764	
	Range	0.3414	Variance	0.004218133554	
	Interquartile range (IQR)	0.0752	Monotocity	Not monotonic	

We also checked Pearson's r correlation for all variables included and not included.

**Fig 1: Pearson's r correlation heatmap**

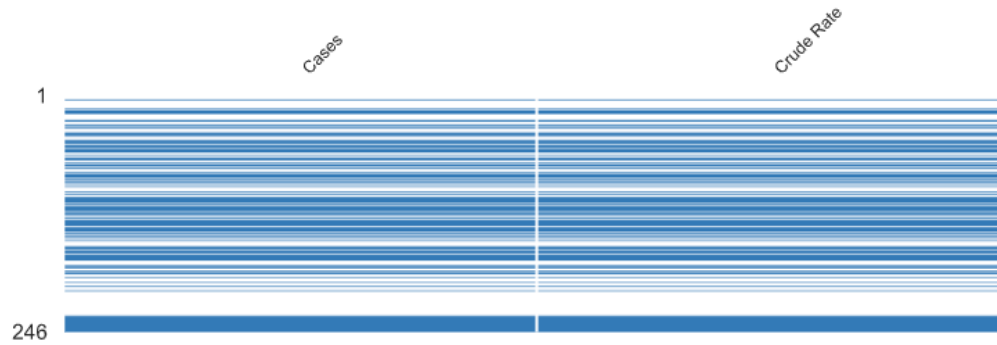


### Data imputation:

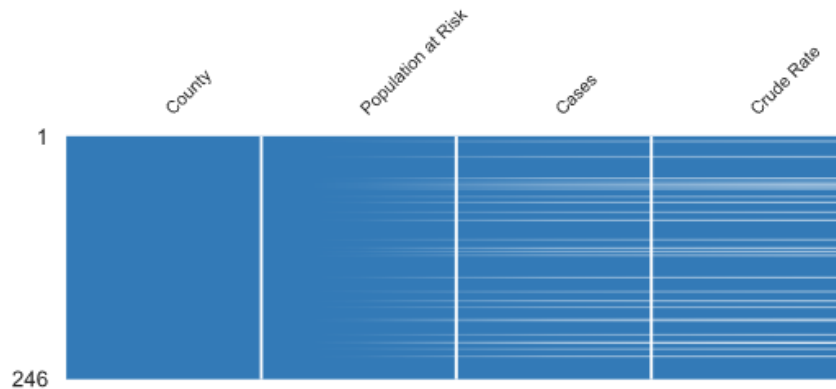
To protect patient identity, if the number of cases is less than 16, then they are suppressed. This causes data suppression for many data. To deal with the issue, we combined data for 10 years. Also, cancer

data does not have strong temporal signal based on our experience. Also, the data was available for only 246 counties

**Fig 2: The nullity matrix below shows the counties with the missing cases for the year 2018.**



**Fig 3: The nullity matrix below shows missing cases after combining data for 10 years.**



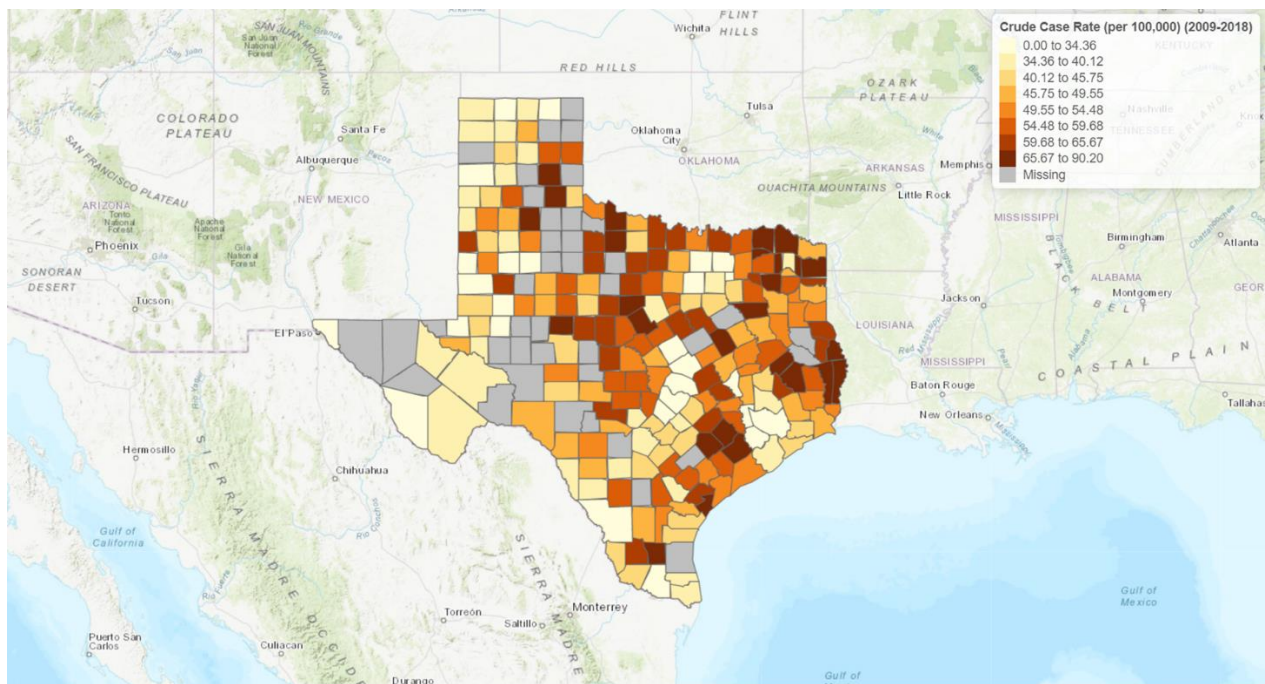
To impute the missing values, we devised our own novel method not used before in any spatial studies. We developed a machine learning model based on Bayesian regression with L2 penalty based on age distribution, expected number of cases, population, and other demographic factors. We ensured that all the imputed values were less than 16. Bayesian ridge estimates a probabilistic model of the regression problem. The prior of the coefficient  $\omega$  is given by a spherical Gaussian. The priors  $\alpha$  and  $\lambda$  are chosen to be gamma distributions, the conjugate prior for the precision of the Gaussian.

$$p(w|\lambda) = \mathcal{N}(w|0, \lambda^{-1}\mathbf{I}_p)$$

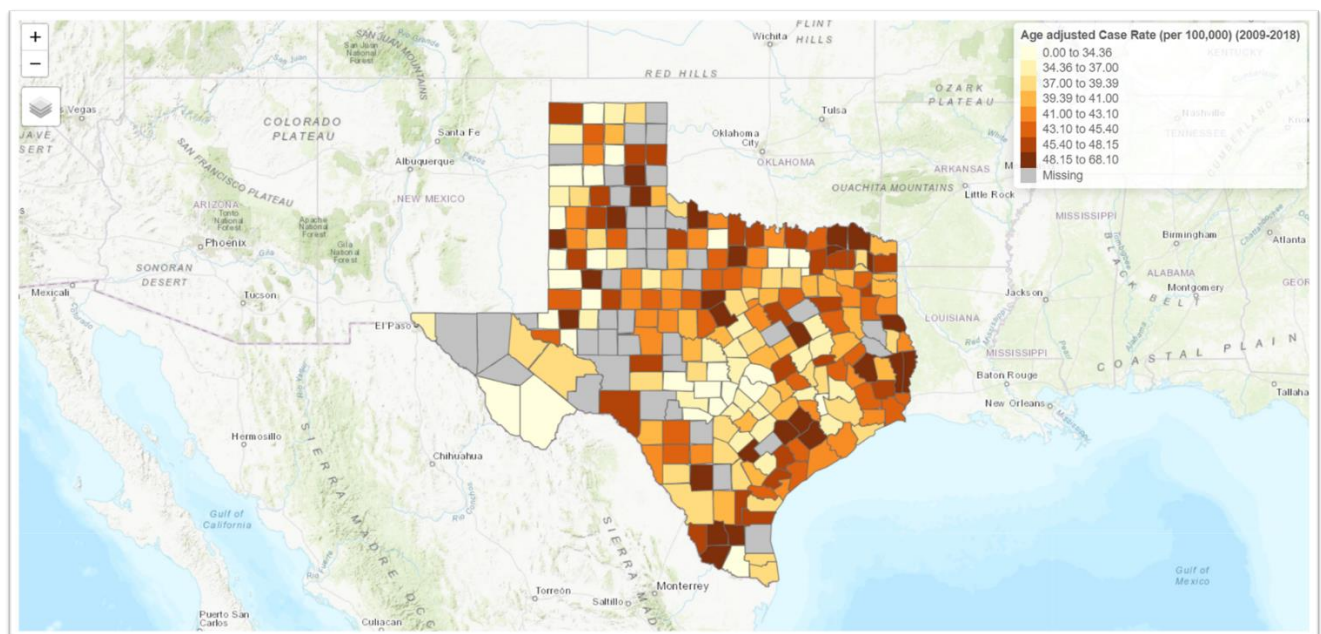
### Exploratory analysis:

We performed exploratory analysis to understand spatial pattern for colorectal cancer incidence. To maintain authenticity, we did not include imputed counts. Fig 4 - 7 below clearly shows that the case rate is associated with population density. The case rate is higher in Eastern and Southern parts of Texas.

**Fig 4: Crude Case Rate in Texas Counties for CRC incidence**

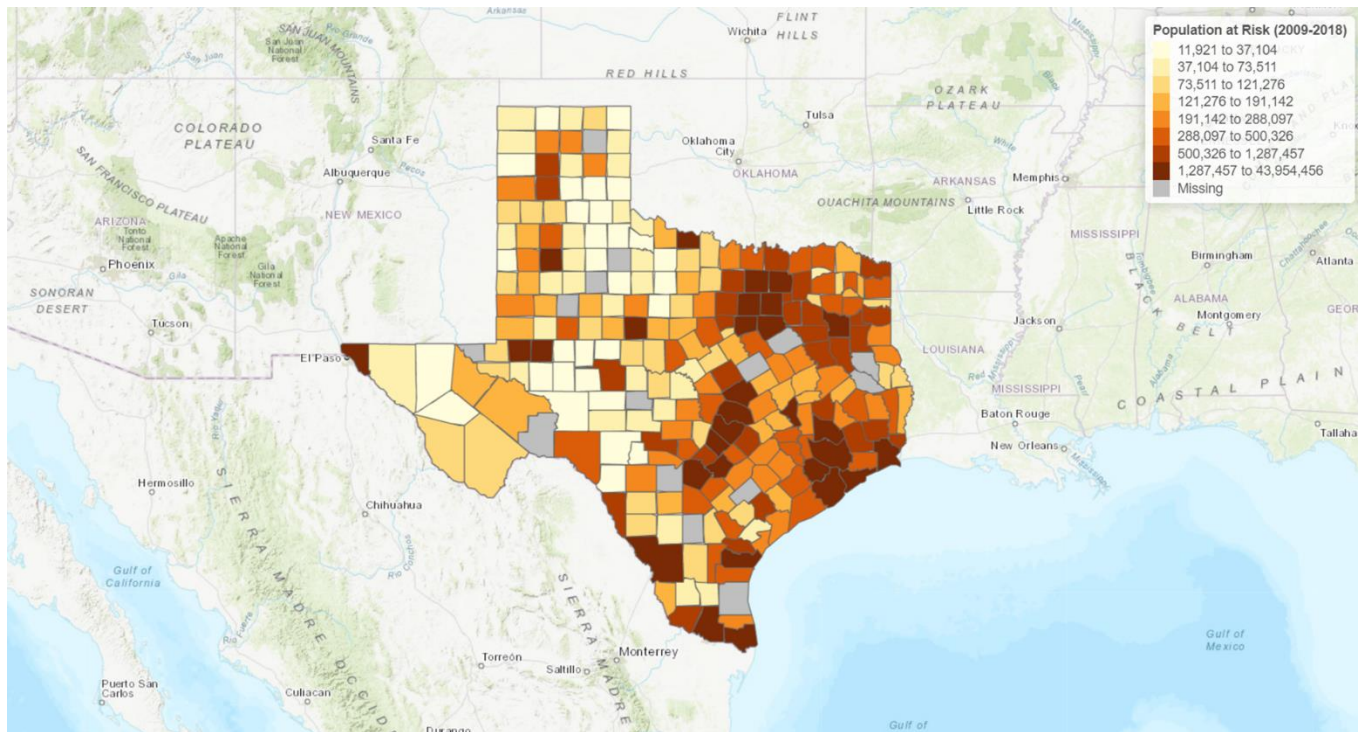


**Fig 5: Age adjusted Case Rate.**

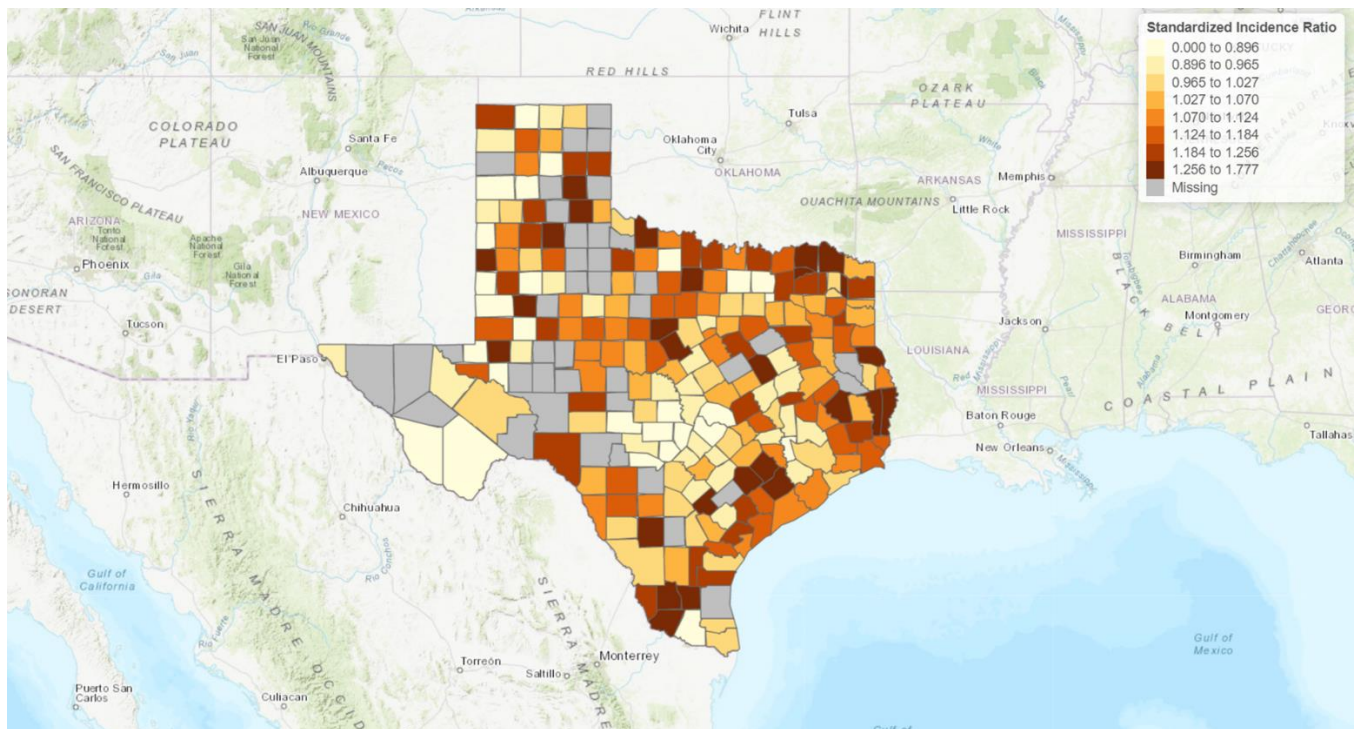




**Fig 6: Population at risk**

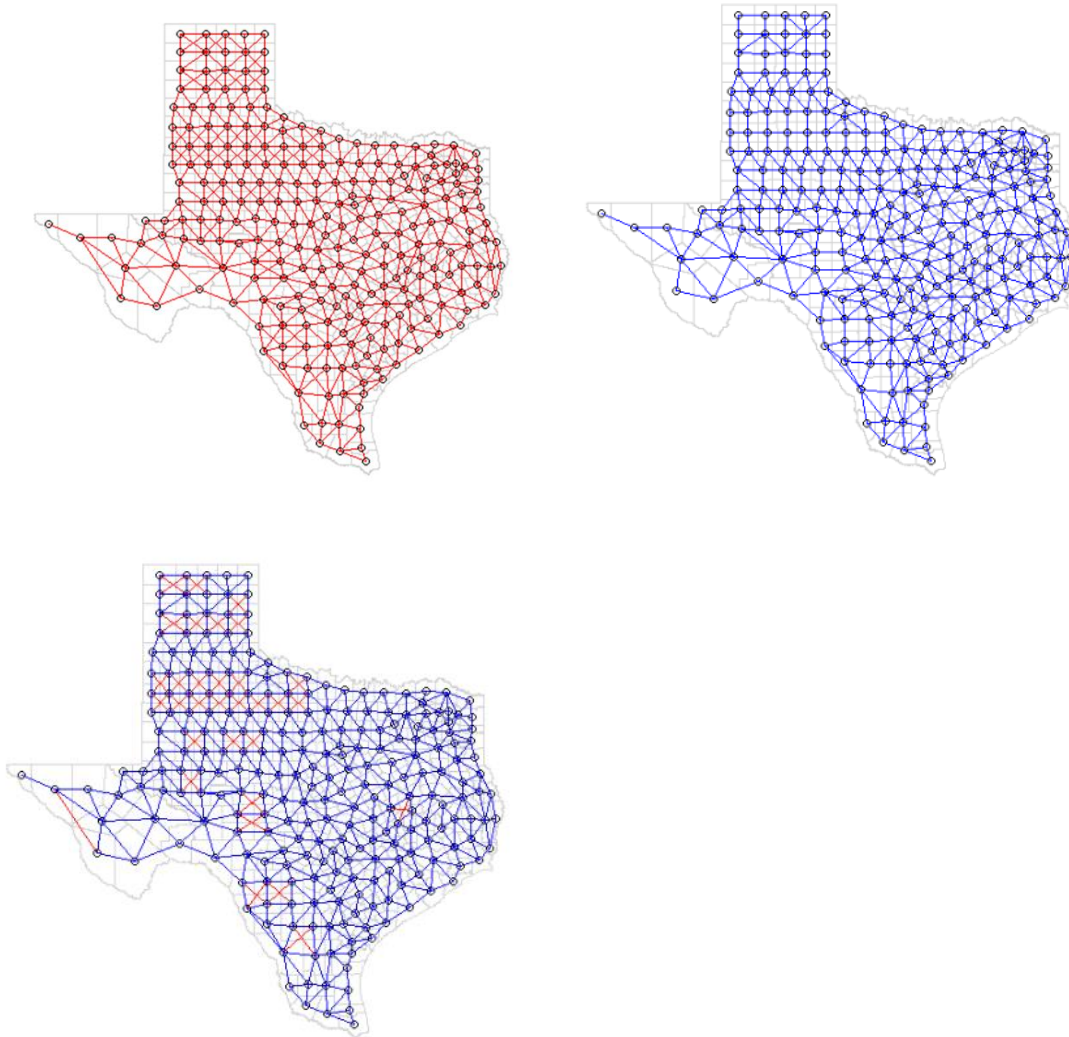


**Fig 7: Standardized Incidence Ratio**



To obtain statistical value for spatial autocorrelation, we measured Moran's I statistic. We used Rook neighbor links, i.e., considered only those counties as neighbors which are touching each other directly.

**Fig 8: Rook vs Queen neighbor links. Red are Queen neighbor links and Blue are Rook neighbor links.**



To understand local spatial autocorrelation, we first created Moran scatterplot. Since the plot is centered on the mean (of zero), all points to the right of the mean have  $z_i > 0$  and all points to the left have  $z_i < 0$ . Values referred to respectively as high and low, in the limited sense of higher or lower than average. We used R package called moran with nb2listw with style=W. nb2listw function supplements a neighbors list with spatial weights where each neighboring polygon will be assigned equal weight (style="W").

To obtain significance value for local spatial autocorrelation, we used LISA (local indicator of spatial association) clusters. It provides a statistic for each location with an assessment of significance. Second, it establishes a proportional relationship between the sum of the local statistics and a corresponding global



statistic. As there are many statistics for global spatial autocorrelation, there will be many corresponding LISA. We have focused on the Local Moran statistic to identify local clusters and local spatial outliers.

## Bayesian analysis of association of county level demographic risk factors with Colorectal Cancer Incidence:

To understand the association of county level demographic variables as described before, we used R-INLA[8], a package in R to run pure Bayesian analyses using integrated nested Laplace approximation and GLM package in R with 4 different models GLM Poisson model, Bayesian Poisson model, Quasi Poisson model, Bayesian Poisson model with non-spatial random effects(iid model with non-spatial random effects), Bayesian spatial model with additional county-level spatial random effect (conditional autoregressive (ICAR) model introduced by Besag (1991))). The level of statistical significance used for this study will be 0.05 and 0.025 and 0.975 quantile will be used for Bayesian models.

## Results

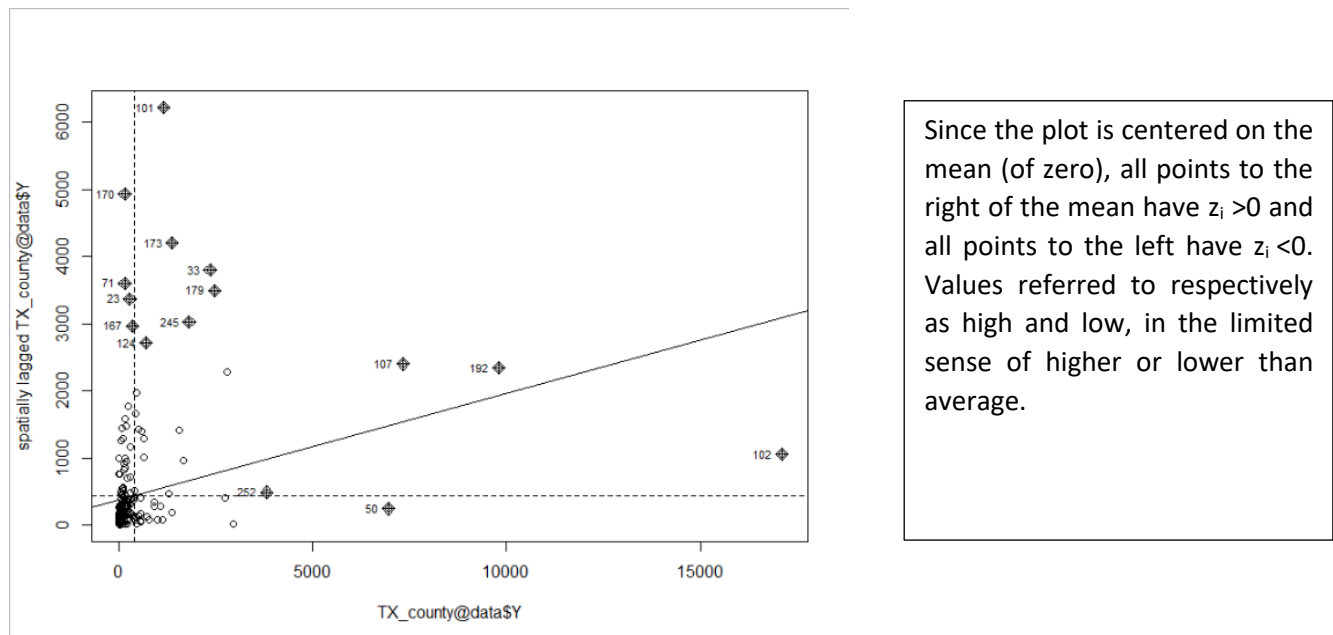
Global Moran's test statistic has correlation score between +1 to -1

+ 1 <- Perfect positive spatial autocorrelation

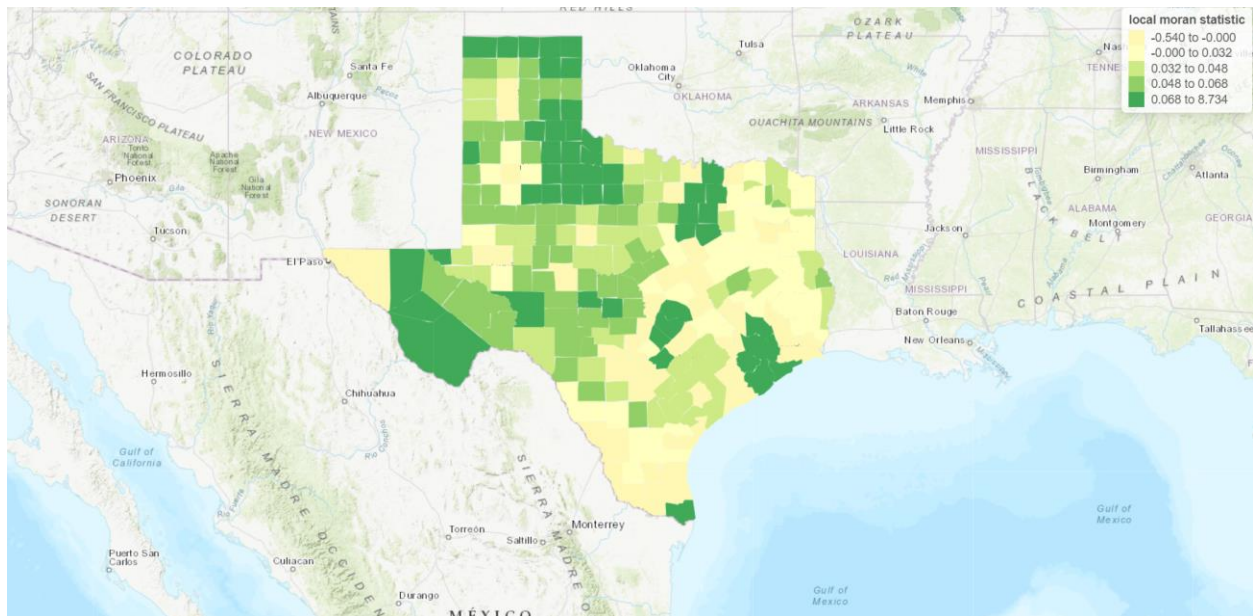
-1 <- Perfect negative spatial autocorrelation

Global Moran I test	Moran I statistic	P value
Randomization	0.16	0.0000003
1000 Monte-Carlo simulation	0.20	0.003

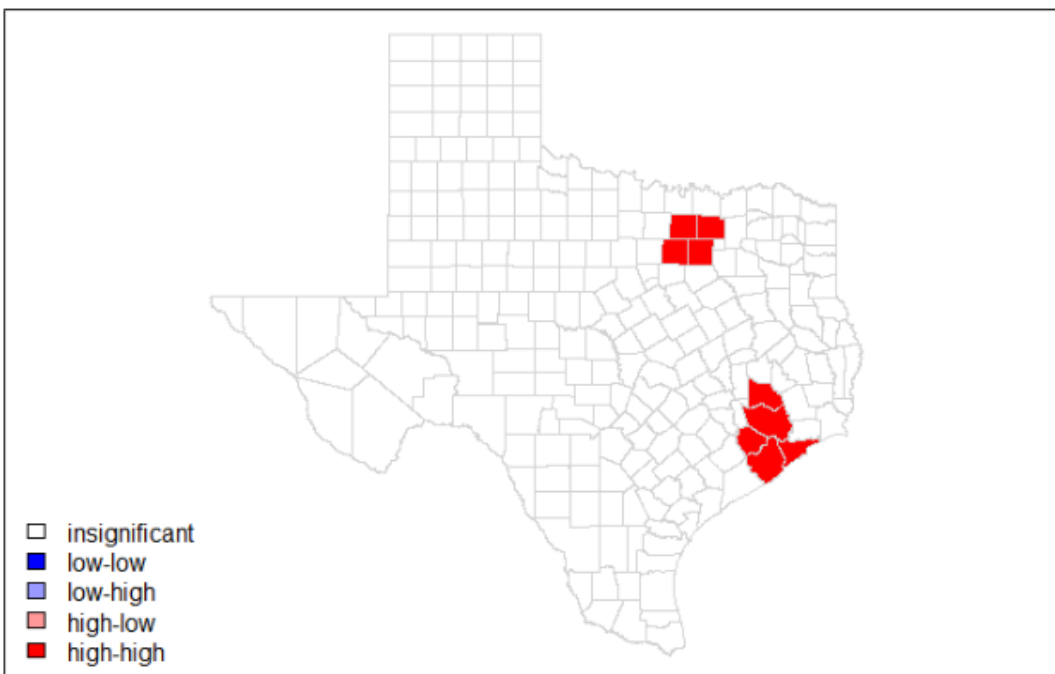
**Fig 9 Moran scatterplot**



**Fig 10: Local Moran statistic mapped to county.**



**Fig 11: Plotting LISA (local indicator of spatial association) clusters**



### Poisson model in INLA: Relative Risk and 95% CI

Features	RR	95% CI
Hispanic	0.946	0.902 – 0.991
White	0.691	0.514 – 0.919
African American	0.822	0.571 – 1.168
Bachelor's degree	0.574	0.478 - 0.685
Smoking	5.249	2.593 – 10.319

### Bayesian non-spatial model: Relative Risk and 95% CI

Features	RR	95% CI
Hispanic	0.984	0.748 – 1.079
White	0.636	0.326 – 1.212
African American	0.660	0.295 – 1.430
Bachelor's degree	0.554	0.383 – 0.794
Smoking	4.993	1.350 – 17.392

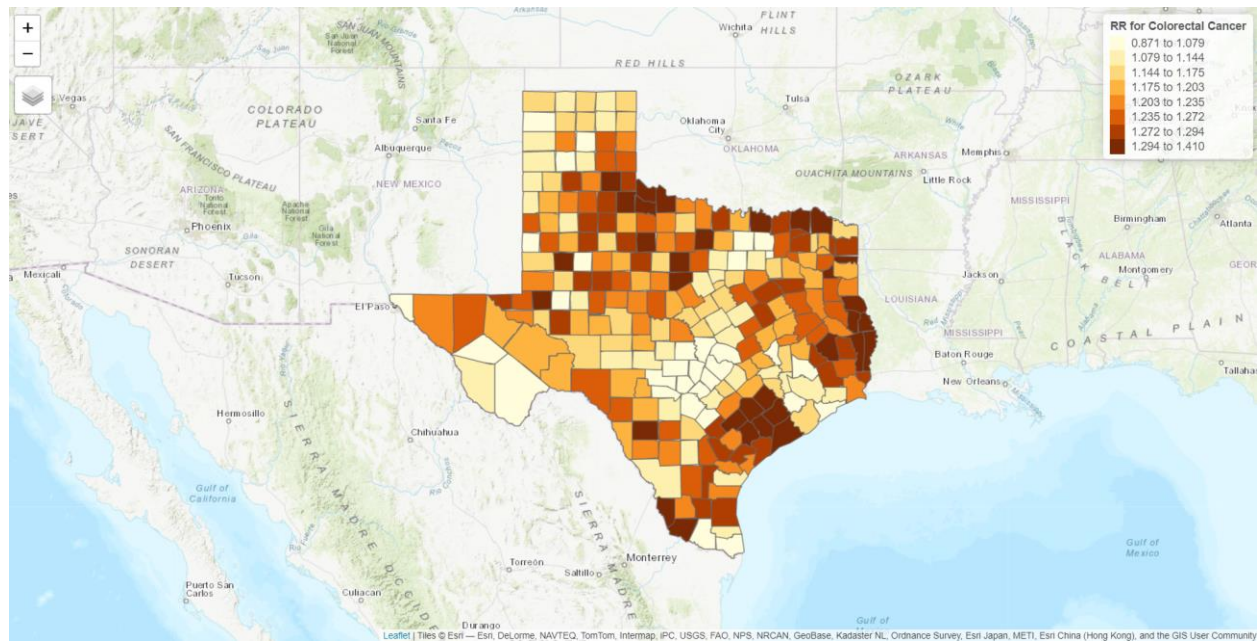
### Bayesian Spatial model: Relative Risk and 95% CI

Features	RR	95% CI
Hispanic	1.059	0.926 – 1.207
White	0.839	0.436 – 1.575
African American	0.896	0.401 – 1.941
Bachelor's degree	0.690	0.474 - 0.993
Smoking	7.614	2 – 27.385

### WAIC values

Model	WAIC
Bayesian Poisson Model	2164
Bayesian non-spatial model (iid model)	2006
Conditional autoregressive (ICAR) model (Bayesian Spatial model)	1992

**Fig 12: Relative Risk for Colorectal Cancer**



## Discussion

Colorectal cancer being the third most common cancer in the United States of America demands significant interest in public health policy making and design intervention. If we can assess spatial clustering and association with demographic variables, it can help understand resource allocation. We started with spatial exploratory analysis. On observing the maps, we can see that the cases are proportionally distributed as per the population density. The counties on the East and the South of Texas show higher incidence rate. To quantify the spatial autocorrelation, we calculated Moran's  $I$  at global and local level. Global Moran's  $I$  was significant and had positive value. This showed positive spatial autocorrelation. We mapped local Moran statistic values. To account for significance in local Moran statistic, we plotted LISA clusters. We see two LISA clusters, one in North East and another in South East region of Texas. These are areas of higher case rate.

To understand association with demographic variables, we built models using GLM and INLA package in R. On comparing WAIC of various models built, we found WAIC lowest for Conditional autoregressive (ICAR) model (Bayesian Spatial model). This shows including spatial random effect improved model performance. Smoking showed highest significant increase in Relative Risk amongst all models. In ICAR model, smoking showed RR of 7.614 (2.0-27.4). We also found that having bachelor's degree reduced CRC incidence thus showing increased CRC screening with education. In ICAR model, bachelor's degree showed RR of 0.690 (0.47-0.99). None of the racial was found to have significant association with CRC incidence.

Our approach has some limitations. First, we imputed suppressed values. The imputed values are not missing but suppressed. Second, we have county level data which has reduced statistical power compared to point level data. Third, we used education level as a proxy to CRC screening awareness based on few studies, but it needs more studies specific to Texas. Third, CRC risk factors are mostly genetic such as



familial adenomatous polyposis (FAP)[9] or hereditary non-polyposis colorectal cancer (Lynch syndrome)[10] and we cannot account that based on demographic variables we had.

However, we believe our approach has provided information about significant clusters and further identified the important association of smoking and education. More resources can be diverted in reducing smoking in counties with higher smoking rate and promoting colorectal cancer awareness in counties with lower education levels.

## Conclusion

Colorectal cancer being one of the most important cancer needs targeted public health campaigns. Through our spatial analysis, we identified clusters having higher incidence of CRC. We also showed positive association of county level smoking and higher education in form of bachelor's degree with incidence of CRC. We also developed a unique, novel statistical imputation method for suppressed cancer data not used before in any other spatial studies. Our study can be further expanded by collecting more demographic variables.

## Competing interests

Authors have no vested interests to declare for. The study was self-funded.

## Authors' contributions

K.P. designed the study, collected the data, ran the analysis, wrote the manuscript. C.B. designed the study, provided the R code, reviewed the analysis, edited the manuscript, and provided feedback and critique for the paper.

## References

1. Colorectal Cancer Statistics | How Common Is Colorectal Cancer?  
<https://www.cancer.org/cancer/colon-rectal-cancer/about/key-statistics.html>. Accessed 2 May 2021.
2. Colorectal Cancer, United States—2007–2016 | CDC. 2020.  
<https://www.cdc.gov/cancer/uscs/about/data-briefs/no16-colorectal-cancer-2007-2016.htm>. Accessed 17 Mar 2021.
3. Doubeni CA, Major JM, Laiyemo AO, Schootman M, Zauber AG, Hollenbeck AR, et al. Contribution of behavioral risk factors and obesity to socioeconomic differences in colorectal cancer incidence. *J Natl Cancer Inst.* 2012;104:1353–62.
4. Rodriguez N, Smith J. The Association Between Education and Colorectal Cancer Screening among United States Veterans Aged 50-75 Years Old: 286. *Off J Am Coll Gastroenterol ACG.* 2016;111:S134.
5. Crookes DM, Njoku O, Rodriguez MC, Mendez EI, Jandorf L. Promoting colorectal cancer screening through group education in community-based settings. *J Cancer Educ Off J Am Assoc Cancer Educ.* 2014;29:296–303.
6. Cancer-Rates.info | Texas. <https://www.cancer-rates.info/tx/>. Accessed 2 May 2021.

7. TAC. <https://imis.county.org/iMIS/CountyInformationProgram>. Accessed 2 May 2021.
8. Rue H, Riebler A, Sørbye SH, Illian JB, Simpson DP, Lindgren FK. Bayesian Computing with INLA: A Review. ArXiv160400860 Stat. 2016. <http://arxiv.org/abs/1604.00860>. Accessed 2 May 2021.
9. Hereditary and Familial Colon Cancer. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3057468/>. Accessed 2 May 2021.
10. Lynch H, Lynch P, Lanspa S, Snyder C, Lynch J, Boland C. Review of the Lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications. Clin Genet. 2009;76:1–18.