Identifying pediatric cancer clusters in Florida using spatially penalized loglinear models

Hao Wang * and Abel Rodríguez †

Abstract

We discuss the identification of pediatric cancer clusters in Florida between 2000 and 2010 using a penalized generalized linear model. More specifically, we introduce a Poisson model for the observed number of cases on each of Florida’s ZIP Code Tabulation Areas (ZCTA) and regularize the associated disease rate estimates using a fused Lasso penalty. Our analysis suggests the presence of a number of pediatric cancer clusters during the period over study, with the largest ones being located around the cities of Jacksonville, Miami, Cape Coral/Fort Meyers and Palm Beach.

1 Introduction

In Amin et al. (2010), an analysis of pediatric cancer records collected in Florida between 2000 and 2007 identified two possible cancer clusters (one in south Florida and one in northeastern Florida) using the SaTScan™ software. In this paper we provide an alternative analysis of an updated version of this dataset covering the years between 2000 and 2010. Our approach relies on a penalized generalized linear model (pGLM) for de novo identification of cancer clusters.

The National Cancer Institute defines a disease cluster as “the occurrence of a greater than expected number of cases of a particular disease within a group of people, a geographic area, or
a period of time”. Although this definition highlights that the notion of disease cluster is purely statistical, it provides little guidance about how to identify them. Accordingly, a number of approaches have been proposed in the literature, examples include methods for point process data such as methods based on distances among points (e.g., see Whittemore et al., 1987, Besag & Newell, 1991 and Tango, 1995) and scan statistics (e.g., see Weinstock, 1981, Openshaw et al., 1987, Kulldorff & Nagarwalla, 1995 and Tango & Takahashi, 2005), as well as methods for aggregated data based on tests of proportions (e.g., see Pothoff & Whittinghill, 1966 and Pothoff & Whittinghill, 1966) or models for spatially correlated data (e.g., see Knorr-Held & Rässer, 2000).

As discussed in the introductory paper of this issue, the data we analyze consists of 6,558 cases of pediatric cancer occurring in Florida between January 2000 and December 2010. Covariates available for each of the patients include race (three categories: African American, Caucasian, and Other), age (four categories: 0-4, 5-9, 10-14, and 15-19 years of age), and sex (binary: female or male). Also, cases are geolocated according to the ZIP Code Tabulation Areas (ZCTAs) of residence of the patient. Hence the focus of this paper is on techniques that allow us to identify disease clusters on data that has been aggregated over space and/or time. Correspondingly, we propose to model the observed number of cases on each of Florida’s ZCTAs using a Poisson model in which over-dispersion in the data is captured by introducing ZCTA-specific random effects, which are regularized (or, alternatively, given a prior distribution) through a fused Lasso penalty (Tibshirani et al., 2005; Friedman et al., 2007; Rinaldo et al., 2009).

Although the data is (at least in principle) spatio-temporal in nature, we aggregate the data on each ZCTA over time and ignore the temporal component. We take this approach because annual counts on individual ZCTAs tend to be very small and because environmental factors affecting cancer incidence rates are likely to operate over long time scales, making inter-annual fluctuations less important than spatial trends. On the other hand, we focus on a fused lasso penalty rather than a more traditional Gaussian conditional autoregressive prior widely used in spatial statistics and disease mapping because the fused lasso induces sparsity in the point estimates generated by the model, allowing us to treat the hypothesis testing problem as an estimation problem. One-dimensional versions of this model have been used in change-point and hot-spot estimation in genomics (e.g., see Tibshirani & Wang, 2008) but, to the best of our knowledge, fused lasso penalties have never been used in the context of disease cluster identification.

The remaining of the paper is organized as follows: Section 2 describes our model for cancer cluster detection and discusses some of its properties. Section 3 describes our computational approach to fitting the model, which relies on non-trivial optimization algorithms. Section 4 presents...
our results for the Florida dataset. Finally, Section 5 discusses some shortcomings of the models, as well as some implications of the results for cancer surveillance in Florida.

## 2 Identification of cancer clusters in Florida using a penalized generalized linear model

In this section we describe the statistical models we use to identify cancer clusters in Florida. We start by considering a model in which we ignore the effect of covariates and discuss modeling the (internally standardized) relative risks for each of the ZCTAs with non-zero pediatric population over the whole period over study. We then explain how these models are extended to account for covariates.

We start by discussing some notation. Let $y_i$ and $n_i$ be, respectively, the total observed number of pediatric cancer cases and the total pediatric population on ZCTA $i$, where $i = 1, \ldots, 979$. The overall disease rate $\bar{\theta}$ is then simply $\bar{\theta} = \frac{\sum_{i=1}^{979} y_i}{\sum_{i=1}^{979} n_i}$. We model the total observed number of cases $y_i$ as a Poisson random variable with intensity $\eta_i$, $y_i \mid \eta_i \sim \text{Poi}(\eta_i)$, where $\log \eta_i = \log n_i + \log \bar{\theta} + \phi_i$ and $\phi_i$ is a random effect (or, alternatively, a frailty term) that captures overdispersion in the data. The value of $\theta_i = \exp\{\phi_i\}$ represents the excess risk in ZCTA $i$, so that $\theta_i > 1$ (or, equivalently, $\phi_i > 0$) suggest areas of increased risk. The log-likelihood associated with this model can be written as

$$
\ell(\phi_1, \ldots, \phi_{979}; y_1, \ldots, y_{979}) = \sum_{i=1}^{979} \left( y_i \left\{ \log n_i + \log \bar{\theta} + \phi_i \right\} - n_i \bar{\theta} \exp\{\phi_i\} \right).
$$

Direct maximization of (1) leads to the trivial estimate $\hat{\theta}_i^{\text{MLE}} = y_i / (n_i \bar{\theta})$. Instead, we propose to maximize a penalized log-likelihood

$$
l_{FL}(\phi_1, \ldots, \phi_{979}; y_1, \ldots, y_{979}) = \ell(\phi_1, \ldots, \phi_{979}; y_1, \ldots, y_{979}) + J_{\lambda, \gamma}(\phi_1, \ldots, \phi_{979}).
$$

The term $J_{\lambda, \gamma}(\phi_1, \ldots, \phi_{979})$ corresponds to a fused lasso penalty (Tibshirani et al., 2005; Friedman et al., 2007; Rinaldo et al., 2009),

$$
J_{\lambda, \gamma}(\phi_1, \ldots, \phi_{979}) = -\lambda \gamma \sum_{i=1}^{I} |\phi_i| - \lambda \sum_{i \sim j} |\phi_i - \phi_j|,
$$

Florida has a total of 983 ZCTAs, but four of them had no pediatric population (and, of course, no cases) during the period we study. Hence, our analysis involves data from only 979 ZCTAs.
where \( \sum_{i' \sim i} \) denotes the sum over all pairs of Florida’s ZCTAs that share a common boundary with each other. An important characteristic of the penalized estimates

\[
\left( \tilde{\phi}_1(\lambda, \gamma), \ldots, \tilde{\phi}_{979}(\lambda, \gamma) \right) = \arg\max_{(\phi_1, \ldots, \phi_{979})} \left\{ \sum_{i=1}^{979} \left( y_i \{ \log n_i + \log \bar{\theta} + \phi_i \} - n_i \bar{\theta} \exp \{ \phi_i \} \right) - \lambda \sum_{i=1}^{979} |\phi_i| - \lambda \sum_{i' \sim i} |\phi_i - \phi_j| \right\}
\]

is that the value for groups of adjacent coefficients can be identical to each other and/or be exactly zero, leading to both to a segmentation of the state into groups of neighboring ZCTAs, and to a classification of those groups as having or not having a relatively risk significantly different from one (and for those groups of ZTCA that have a relative risk significantly different from one, a shrunk estimate of the corresponding relatively risk). Maximizing the penalized log-likelihood allows us to treat the problem of simultaneously testing multiple hypotheses as an estimation problem that can be efficiently solved (see Section 3). We exploit this property to define \( k \)-th cancer cluster in the sample as a group of adjacent ZCTAs, with positive log-relative risk, i.e., as a group of indexes \( i_1, \ldots, i_{m_k} \) such that for every \( j = 1, \ldots, m_k \) we have \( \tilde{\phi}_{i_j}(\lambda, \gamma) > 0 \) and for some \( j' = 1, \ldots, m_k \) we have \( i_j \sim i_{j'} \).

From a Bayesian perspective, the fused lasso penalty can be motivated as corresponding to a (proper) prior of the form,

\[
p(\phi_1, \ldots, \phi_{979} | \lambda, \gamma) = \frac{1}{C(\lambda, \gamma)} \exp \left\{ -\lambda \gamma \sum_{i=1}^{979} |\phi_i| - \lambda \sum_{i' \sim i} |\phi_i - \phi_j| \right\},
\]

where \( C(\lambda, \gamma) = \int \left\{ -\lambda \gamma \sum_{i=1}^{979} |\phi_i| - \lambda \sum_{i' \sim i} |\phi_i - \phi_j| \right\} d\phi \) is the normalizing constant. This prior can be written as a mixture of Gaussian distributions (e.g., see Kyung et al., 2010) and, for the case of \( \gamma = 0 \), corresponds to the marginal prior induced by a hierarchical Gaussian conditional autoregressive model (CAR) model (e.g., see Rodriguez & Mendoza, 2014). More generally, the penalty parameter \( \lambda \) controls the level of similarity in the estimates for neighboring regions; \( \lambda = 0 \) implies that the frailty terms are independent a priori and the overdispersion in the counts does not follow any spatial pattern, while \( \lambda \to \infty \) leads to a model in which the level of overdispersion is the same in all ZCTAs.

The performance of the model depends critically on the value of the penalty parameters \( \lambda \) and \( \gamma \), which need to be estimated from the data. In this paper we use Akaike’s information criterion.
(AIC) (Akaike, 1974),

\[
AIC(\lambda, \gamma) = -2 \log l \left( \tilde{\phi}_1(\lambda, \gamma), \ldots, \tilde{\phi}_{979}(\lambda, \gamma); y_1, \ldots, y_{979} \right) + 2\psi \left( \tilde{\phi}_1(\lambda, \gamma), \ldots, \tilde{\phi}_{979}(\lambda, \gamma) \right),
\]

where \(\psi \left( \tilde{\phi}_1(\lambda, \gamma), \ldots, \tilde{\phi}_{979}(\lambda, \gamma) \right)\) represents the equivalent number of parameters associated with \(\lambda\) and \(\gamma\) (in this case, the number non-zero blocks of coefficients that are obtained when the values of \(\lambda\) and \(\gamma\) are used to computed the penalized estimates in (4), see Zou et al., 2007 and Tibshirani et al., 2012).

A similar formulation can be used to account for the effect of the available covariates (age, race and sex). In particular, let \(y_{i,j,k,l}\) and \(n_{i,j,k,l}\) correspond to the number of pediatric cancer cases and the total pediatric population in ZCTA \(i = 1, \ldots, 979\), age group \(j = 1, \ldots, 4\), race \(k = 1, \ldots, 3\) and sex \(l = 1, 2\) and define the average incidence rate for each of these subpopulations as \(\bar{\theta}_{j,k,l} = \frac{\sum_{i=1}^{979} y_{i,j,k,l}}{\sum_{i=1}^{979} n_{i,j,k,l}}\). We model the counts \(y_{i,j,k,l}\) by assuming the excess risk in ZCTA \(i\) is the same for all subpopulations, i.e., we let \(y_{i,j,k,l} \mid \eta_{i,j,k,l} \sim \text{Poi}(\eta_{i,j,k,l})\) where \(\log \eta_{i,j,k,l} = \log n_{i,j,k,l} + \log \bar{\theta}_{j,k,l} + \phi_i\). Under this formulation, covariate-adjusted estimates of the excess risk can be obtained by solving

\[
\left( \tilde{\phi}_1(\lambda, \gamma), \ldots, \tilde{\phi}_{979}(\lambda, \gamma) \right) = \arg\max_{(\phi_1, \ldots, \phi_{979})} \left\{ \sum_{i=1}^{979} \sum_{j=1}^{4} \sum_{k=1}^{3} \sum_{l=1}^{2} \left( y_{i,j,k,l} \left\{ \log n_{i,j,k,l} + \log \bar{\theta}_{j,k,l} + \phi_i \right\} - n_{i,j,k,l} \bar{\theta}_{j,k,l} \exp\{\phi_i\} \right) \right. \\
- \lambda \gamma \sum_{i=1}^{I} |\phi_i| - \lambda \sum_{i' \sim i} |\phi_i - \phi_{i'}| \right\}. \quad (6)
\]

## 3 Computational implementation

We solve the maximization problems in (4) and (6) using a variation of the “split-Bregman” algorithm discussed in Goldstein & Osher (2009). The algorithm is iterative and relies on a second-order Taylor approximation to the Poisson likelihood and on the introduction of two auxiliary vectors \(u\) and \(d\) that allow us to break the optimization problem into coupled subproblems that are, individually, easy to solve. In the case of (4), the algorithm takes the following form.

1. Initialize \(\hat{\phi}^{(0)}\) and pick a tuning parameter \(\xi\) that controls the rate of convergence of the algorithm.
2. Starting with \( k = 0 \) and until convergence, repeat:

(a) Update the parameters of the quadratic approximation to the likelihood function by setting

\[
H^{(k)} = - \text{diag} \left\{ n_1 \tilde{\theta} \exp \left\{ \tilde{\phi}_1^{(k)} \right\}, \ldots, n_I \tilde{\theta} \exp \left\{ \tilde{\phi}_I^{(k)} \right\} \right\}
\]

and the vector

\[
h^{(k)} = \begin{pmatrix} y_1 - n_1 \tilde{\theta} \exp \left\{ \tilde{\phi}_1^{(k)} \right\} \\ \vdots \\ y_I - n_I \tilde{\theta} \exp \left\{ \tilde{\phi}_I^{(k)} \right\} \end{pmatrix}.
\]

(b) Initialize \( \tilde{\phi}^{(k+1,0)} = \tilde{\phi}^{(k)} \), \( u^{(k+1,0)} \) and \( d^{(k+1,0)} \).

(c) Starting with \( l = 0 \) and until convergence, repeat:

(i) Update,

\[
A^{(k,l)} = \xi \lambda L^T L - H^{(k)}
\]

and

\[
a^{(k,l)} = \xi \lambda L^T \left( u^{(k+1,l)} - d^{(k+1,l)} \right) + h^{(k)} - H^{(k)} \tilde{\phi}^{(k)}.
\]

(ii) Initialize \( \tilde{\phi}^{(k+1,l,0)} = \tilde{\phi}^{(k+1,l)} \).

(iii) Starting with \( s = 0 \) and until convergence, iterate the following step:

\[
\tilde{\phi}_i^{(k+1,l,s+1)} = S \left( \alpha_i^{(k,l)} - \sum_{j<i} A_{i,j}^{(k,l)} \frac{\tilde{\phi}_j^{(k+1,l,s+1)}}{A_{i,i}^{(k,l)}} - \sum_{j>i} A_{i,j}^{(k,l)} \frac{\tilde{\phi}_j^{(k+1,l,s)}}{A_{i,i}^{(k,l)}} - \frac{\gamma \lambda}{A_{i,i}^{(k,l)}} \right),
\]

where \( S(x, \delta) = \text{sgn}(x) \max\{0, |x| - \delta\} \) is the soft thresholding operator.

(iv) When the previous sub-iterations have converged, set \( \tilde{\phi}^{(k+1,l+1)} = \tilde{\phi}^{(k+1,l,\infty)} \).

(v) Set \( u^{(k+1,l+1)} = S \left( d^{(k+1,l)} + \lambda L \tilde{\phi}^{(k+1,l+1)} \right), \frac{1}{\xi} \), where the thresholding operator is applied componentwise.

(vi) Set \( d^{(k+1,l+1)} = d^{(k+1,l)} + \lambda L \tilde{\phi}^{(k+1,l+1)} - u^{(k+1,l+1)} \).
(d) Once this sub iteration has converged, set \( \tilde{\phi}^{(k+1)} = \tilde{\phi}^{(k+1, \infty)} \)

3. Once this iterations have converged, report \( \tilde{\phi}^{(\infty)} \) as your point estimate for \( \phi \).

We present details of the derivation of this algorithm in Appendix A, and note that generalizing it for maximizing (6) is straightforward. (The only difference is in the structure of \( H \) and \( h \).) We implemented the iterative thresholding in step (iii) above using the function `crossProdLasso` from the \texttt{R} package \texttt{scout} (Witten & Tibshirani, 2011). All sub iterations were considered to have converged when the relative \( L^2 \) error in the estimate of the vector \( \phi \) was less than \( 10^{-4} \). For the tuning parameter \( \xi \), we explored values between 4 and 40 and found that the performance of the algorithm was quite robust to this choice.

To explore the effect of initial values, we initialize the Poisson generalized lasso algorithm at two different and meaningful values. The first is \( \tilde{\phi}^{(0)} = 0 \), corresponding to zero log relative risks for all regions, and the second is \( \tilde{\phi}^{(0)} = \{ \log(y_i/n_i) - \log(\bar{\theta}) \} \), corresponding to the observed log relative risks. Because some regions have zero incidents, implying a value of \(-\infty\) for the observed log relative risk, we set the initial log relative risk in these regions equal to the smallest finite value in the sample. The algorithm seems to be robust to the choice of initial values, as the results agree for up to the four decimal place.

To select the tuning parameters \( \lambda \) and \( \gamma \) we evaluate \( AIC(\lambda, \gamma) \) over a grid of values of \( (\gamma, \lambda) \). In particular we take \( \gamma \in \{0.5, 1, 2\} \), indicating the ratio of the strength of the pure lasso penalty over that of the fusion penalty is in the range of 50\% and 200\%. For \( \lambda \), we first run the path algorithm of Tibshirani et al. (2012) for solving the least square generalized lasso approximation (8) at \( \tilde{\phi} = 0 \) and then use the output values of \( \lambda \), at which the solution path changes slope, as the grid for the Poisson generalized lasso.

**4 Results**

**4.1 Relative risks without adjusting for covariates**

Figure 4.1 shows the raw and estimated overall relative risks for each of Florida’s ZCTAs before adjusting for race, gender or ethnicity. These point estimates were generated using the optimal values of \( \tilde{\gamma} = 1 \) and \( \tilde{\lambda} = 0.718 \) obtained using AIC. By comparing these two maps we note that our algorithm has the desired effect of smoothing out the raw observations, leading to estimates that involve a large number of ZCTAs with no increased or reduced relative risks. Our approach
also identified 26 possible clusters with elevated overall relative risk involving 274 ZCTAs; some of these clusters had raw risks that were up to 4 times higher as Florida’s average. Many of these clusters (19 out of the 26) correspond to either isolated ZCTAs or small clusters with only two or three ZTCAs in them. However, the largest clusters (with 91, 73, 32 and 24 ZCTAs and an average at-risk pediatric population of 341,755, 579,902, 120,241 and 162,272 individuals each year) are located in north Florida (the Jacksonville metro area and counties to the West), the Miami metro area, the Cape Coral-Fort Myers metro area and counties to the East, and the county of Palm Beach. The clusters we identify mostly fall within the boundaries of the clusters identified in Figure 1 of Amin et al. (2010); in particular, we seem to find the same small cluster in central Florida that the aforementioned authors identified in their original dataset. However, our clusters tend to be much smaller, suggesting that our methodology allows for more precise identification.

Figure 1: Raw (left) and estimated (right) overall log relative risks for pediatric cancers in Florida.

Table 1 presents the raw incidence rates of pediatric cancer in different covariate-driven subgroups for these four large clusters, and compares them again the average incidence rate in Florida for the same groups. Note that raw incidence rates in these clusters are between 33% and 39% higher than for Florida as a whole, which is substantial. Also, although the specific patterns vary in the different clusters, disease rates are elevated in almost every subgroup. The main (and some-
Table 1: Raw incidence rates of pediatric cancer in different covariate-driven subgroups for the four largest clusters identified by our model when there is no adjustment for covariates.

what surprising) exception is the racial group “Other” (made of mostly Hispanics) in the Cape Coral-Fort Myers region, which has a very low incidence rate compared to the Florida average for this same group. However, since the at-risk population is small for this subgroup this might just be the product of chance.

Finally, in order to explore the temporal structure of the data, we present in Figure 2 time series plots of the annual (raw) incidence rate for the four most important clusters we identified and compare them against the overall incidence in Florida. The rates for the North Florida and the Miami clusters are consistently above those of Florida for the whole 11 years of data under analysis. The rates for the Cape Coral and the Palm Beach County clusters also tend to be above those of Florida as a whole, with a couple of exceptions. In the case of the Cape Coral cluster, a spike in cases arose in 2002, followed by a big drop in 2004 and then a recovery towards the average level. On the other hand, for the Palm Beach county cluster we observe an important drop in the number of cases in 2005 followed by an important increase in 2008.

### 4.2 Relative covariate-adjusted risks

Figure 3 presents estimates of the relative risks computed under the covariate-adjusted model in (6). For this map, the optimal values for the smoothing parameters are $\hat{\gamma} = 1$ and $\hat{\lambda} = 0.717$, essentially identical to those in Section 4.1. Note that this map is very similar to the one presented in the right panel of Figure 4.1. We now detect 24 possible clusters involving a total of 276 ZCTAs. Table 2 and appendix B present a more detailed comparison of the four major clusters under each of the models; note that there is substantial agreement between them, with the cluster under one model being almost completely a subset of the respective cluster under the other.

Similarly to Table 1, Table 3 shows the raw incidence rates of pediatric cancer in different covariate-driven subgroups for the four largest clusters identified by our second model, and com-
Figure 2: Time series plots of the annual incidence rate for four of the clusters (North Florida, Miami, Cape Coral and Palm Beach) compared against that in Florida.

pares them again the average incidence rate in Florida for the same groups. As would be expected the results are very similar to those in Section 4.1, with the Miami cluster still exhibiting a particularly large incidence rate among members of the “Other” racial group, and Cape Coral showing a particularly small incidence rate for the same group.

### 4.3 Validation

Since the small-sample properties of the fused lasso as a model selection mechanism are not well understood, we validate our results by undertaking two small simulation studies to assess the prob-

<table>
<thead>
<tr>
<th></th>
<th>Not adjusted for covariates</th>
<th>After adjusting for covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of ZCTAs</td>
<td>Total pediatric population</td>
</tr>
<tr>
<td>North Florida</td>
<td>91</td>
<td>341 755</td>
</tr>
<tr>
<td>Miami</td>
<td>73</td>
<td>579 902</td>
</tr>
<tr>
<td>Cape Coral</td>
<td>32</td>
<td>120 241</td>
</tr>
<tr>
<td>Palm Beach</td>
<td>24</td>
<td>162 272</td>
</tr>
</tbody>
</table>

Table 2: Characteristics of the four major clusters under our two models.
Table 3: Raw incidence rates of pediatric cancer in different covariate-driven subgroups for the four largest clusters identified by our model after adjusting for covariates.

<table>
<thead>
<tr>
<th></th>
<th>Fitted Overall</th>
<th>Raw Overall</th>
<th>Incidence (per 100,000 children per year)</th>
<th>Sex</th>
<th>Race</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>0-4</td>
<td>5-9</td>
<td>10-14</td>
</tr>
<tr>
<td>North Florida</td>
<td>15.6</td>
<td>18.9</td>
<td>28.9</td>
<td>14.1</td>
<td>15.1</td>
</tr>
<tr>
<td>Miami</td>
<td>16.5</td>
<td>18.0</td>
<td>28.4</td>
<td>13.9</td>
<td>14.9</td>
</tr>
<tr>
<td>Palm Beach</td>
<td>15.0</td>
<td>18.0</td>
<td>28.2</td>
<td>14.5</td>
<td>16.2</td>
</tr>
<tr>
<td>Cape Coral</td>
<td>15.7</td>
<td>19.1</td>
<td>31.3</td>
<td>14.5</td>
<td>16.2</td>
</tr>
<tr>
<td>Florida</td>
<td>–</td>
<td>13.6</td>
<td>21.8</td>
<td>11.1</td>
<td>11.5</td>
</tr>
</tbody>
</table>

Figure 3: Covariate-adjusted log relative risks.
<table>
<thead>
<tr>
<th>Method</th>
<th>Proportion of simulations in which clusters were identified</th>
<th>Average proportion of regions identified as part of a cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>pGLM</td>
<td>0.79</td>
<td>0.045 (0.056)</td>
</tr>
<tr>
<td>BN (K=20)</td>
<td>0.97</td>
<td>0.005 (0.003)</td>
</tr>
<tr>
<td>BN (K=200)</td>
<td>0.98</td>
<td>0.020 (0.016)</td>
</tr>
<tr>
<td>KN (R = 0.02)</td>
<td>1.00</td>
<td>0.039 (0.014)</td>
</tr>
<tr>
<td>KN (R = 0.2)</td>
<td>1.00</td>
<td>0.04 (0.033)</td>
</tr>
</tbody>
</table>

Table 4: Performance of three different algorithms under our first simulation scenario. Standard errors are in parentheses.

Our first simulation study is carried out under a model in which there are no cancer clusters in Florida. More specifically, we generate 100 datasets $y_1^*, \ldots, y_{100}^*$, so that the number of cases in ZCTA $i = 1, \ldots, 979$ for dataset $m = 1, \ldots, 100$ is given by $y_{m,i}^* \sim \text{Poi}(n_i \bar{\theta})$, where $\bar{\theta} = \frac{\sum_{i=1}^{979} y_i}{\sum_{i=1}^{979} n_i} = 0.000136$ is the average overall pediatric cancer risk observed in Florida between 2000 and 2010. For each of these samples we computed the number of clusters identified by the model as well as the proportion of regions identified as being part of a cluster by each of the three procedures discussed above (see Table 4 and Figure 4).

Figure 4: Histogram of the number of high relative risk ZCTA ($\hat{\phi}_i > 0$) by our penalized generalized linear model on each of the 100 datasets generated under our first simulation scenario.

Our second simulation study involves 100 datasets $y_1^*, \ldots, y_{100}^*$ where $y_{m,i}^* \sim \text{Poi}(n_i \bar{\theta} \hat{\theta}_i^*), \ldots, y_{100}^*$, where $y_{m,i}^* \sim \text{Poi}(n_i \bar{\theta} \hat{\theta}_i^*)$. 
Table 5: Performance of three different algorithms under our second simulation scenario. Standard errors are in parentheses.

$$\tilde{\theta}_1, \ldots, \tilde{\theta}_{979}$$ correspond to the overall relative risks for the different ZCTAs reported in Section 4.1. Table 5 presents values for the probability of not identifying any cluster in the data, as well as the sensitivity, specificity and Matthews correlation coefficient associated with each of the three methods for this second scenario.

The results presented above suggest that the pGLM has the best overall performance, presenting the highest MCC coefficient and the lowest probability of detecting a false cluster. However, even though our model has the lowest probability of detecting a false cluster, that probability is moderately high, suggesting that the model has a moderately large chance of detecting a spurious cluster.

### 5 Discussion

Our analysis of the Florida data suggests the presence of a number of pediatric cancer clusters, with the largest ones being roughly located around the cities of Jacksonville, Miami, Cape Coral/Fort Meyers and Palm Beach and covering about a quarter of the total pediatric population in the state. We estimate that the risk of pediatric cancers in these regions is at least 30% higher than the statewide average risk. Importantly, these results seemed to be robust to the inclusion of demographic information. However, our validation using a simulation study suggests that these results must be taken with a grain of salt. Indeed, although our approach has higher sensitivity and specificity and lower type I error rate than the algorithms we compared it against, it still tends to incorrectly identify at least one cluster in datasets that have been generated under a Poisson model with a constant rate in at least 79% of the cases. In spite of this somewhat negative result, we believe that at least some of the biggest clusters are indeed real because of the high MCC index in our method and the fact that the number of ZCTAs with elevated relative risks in the Florida data is much larger.
than the numbers we observed when applying our method to data simulated under the null model.

Another potential shortcoming of our approach is that overdispersion is captured only by spatial random effects. Although this type of assumption is common in the literature on disease mapping, it can potentially lead to the detection of clusters even if the overdispersion does not followed a well defined spatial pattern (e.g., if the data arises from a negative binomial distribution). Although the literature on cancer cluster identification is ambivalent about whether all sorts of overdispersion should suggest the presence of clusters or not, we believe that some sort of spatial coherence in the structure of the clusters is desirable. That is the reason why in our discussion of the results we have focused on the four largest clusters identified by our algorithm. A more conservative approach that would deal with this issue would also include independent random effects to account for non-spatial structure in the overdispersion.

A Derivation of the computational algorithm

We derive our algorithm for solving (4) for a slightly more general model where the log relative risk in region $i$ is modeled as a linear function of a set of predictors $x_i \in \mathbb{R}^p$ and we assume a generalized lasso penalty. In this case, the log likelihood function takes the form

$$
\ell(\beta; y) = \sum_{i=1}^{I} y_i (\log n_i + \log \tilde{\theta} + x_i^T \beta) - \tilde{\theta} \sum_{i=1}^{I} n_i \exp \{x_i^T \beta\},
$$

and the fused lasso penalty (3) can also be written in a general way as

$$
J_{\lambda, \gamma}(\beta) = -\lambda \gamma ||\beta||_1 - \lambda ||L \beta||_1.
$$

where $||u||_1 = \sum |u_i|$ denotes the $L^1$ norm of the vector $u$, and $L$ is a pre-specified $m \times p$ penalty matrix. The random effect model in (1)–(3) corresponds to the special case of $x_i = e_i$ where $e_i$ has all entries 0 except that the $i$th entry equals 1, $\beta = (\phi_1, \ldots, \phi_{979})^T$, and $L$ is a (very sparse) pairwise difference matrix whose rows correspond to pairs of ZCTAs that share a common boundary. Similarly, (6) can be written in a similar form by extending the sum over ZCTAs in (7) to also include sums over all demographic groups.

Recall that the (unpenalized) log-likelihood (7) can be optimized using iteratively reweighted least squares (IRLS), i.e., by iteratively computing

$$
\hat{\beta}^{(k+1)} = \arg\max_{\beta} Q(\beta | \hat{\beta}^{(k)}),
$$

14
where \( Q(\beta | \beta^{(k)}) \) is obtained by a second-order expansion of (7) around the previous iterate \( \hat{\beta}^{(k)} \),

\[
Q(\beta | \hat{\beta}^{(k)}) = \left( \beta - \hat{\beta}^{(k)} \right)^T h(\hat{\beta}^{(k)}) + \frac{1}{2} \left( \beta - \hat{\beta}^{(k)} \right)^T H(\hat{\beta}^{(k)}) \left( \beta - \hat{\beta}^{(k)} \right),
\]

with

\[
h(\hat{\beta}^{(k)}) = \frac{\partial}{\partial \beta} \log \ell(\beta; y) \bigg|_{\beta = \hat{\beta}^{(k)}} = \sum_{i=1}^{I} x_i \left( y_i - \bar{\theta}_n \exp \left\{ x_i^T \hat{\beta}^{(k)} \right\} \right),
\]

and

\[
H(\hat{\beta}^{(k)}) = \frac{\partial^2}{\partial \beta^2} \log \ell(\beta; y) \bigg|_{\beta = \hat{\beta}^{(k)}} = -\sum_{i=1}^{I} x_i x_i^T \bar{\theta}_n \exp \left\{ x_i^T \hat{\beta}^{(k)} \right\}.\]

Similarly, we propose to optimize (4) by iteratively solving (e.g., see Friedman et al., 2010 and Krishnapuram et al., 2005)

\[
\hat{\beta}^{(k+1)} = \arg\min_{\beta} \left\{ -Q(\beta | \hat{\beta}^{(k)}) + \gamma \lambda ||\beta||_1 + \lambda ||L_\beta||_1 \right\},
\]

(8)

where each optimization problem in the sequence is accomplished using a variation of the “split-Bregman” algorithm (Goldstein & Osher, 2009) described in the next subsection.

A.1 Solving the fused lasso problem using the “split-Bregman” algorithm

To derive the “split-Bregman” algorithm, introduce a new variable \( u = \lambda L_\beta \), so that the solution to (8) is equivalent to the solution of the following constrained minimization problem,

\[
\left( \beta^{(k+1)}, u^{(k+1)} \right) = \arg\min_{\beta, u} \left\{ -Q(\beta | \hat{\beta}^{(k)}) + \gamma \lambda ||\beta||_1 + ||u||_1 \right\} \quad \text{subject to} \quad u = \lambda L_\beta,
\]

This problem can be solved using an iterative procedure called the Bregman iteration (Bregman, 1967; Osher et al., 2005),

\[
\left( \beta^{(k+1,l+1)}, u^{(k+1,l+1)} \right) = \arg\min_{u, \beta} \left\{ -Q(\beta | \hat{\beta}^{(k)}) + \gamma \lambda ||\beta||_1 + ||u||_1 \right. \nonumber \\
+ \xi \left. \left\| u - \lambda L_\beta - d^{(k+1,l)} \right\|_2^2 \right\},
\]

(9)

\[
d^{(k+1,l+1)} = d^{(k+1,l)} + \lambda L_\beta \beta^{(k+1,l+1)} - u^{(k+1,l+1)},
\]

(10)

where the added \( L^2 \) norm, \( || \cdot ||_2 \), of the vector \( u - \lambda L_\beta - d^{(k+1,l)} \) is used to enforce the constraint \( u = \lambda L_\beta \) and \( \xi \) is a tuning parameter that controls how fast the constraint is enforced. The final
algorithm is obtained by splitting (9) into two separate optimization steps,

\[
\beta^{(k,l+1)} = \arg\min_{\beta} \left\{ -Q(\beta | \hat{\beta}^{(k)}) + \gamma \lambda \|\beta\|_1 + \frac{\xi}{2} \|u^{(k,l)} - \lambda L \beta - d^{(k,l)}\|_2^2 \right\}, \quad (11)
\]

\[
u^{(k,l+1)} = \arg\min_{u} \left\{ \|u\|_1 + \frac{\xi}{2} \|u - \lambda L \beta^{(k,l+1)} - d^{(k,l)}\|_2^2 \right\}, \quad (12)
\]

\[
d^{(k,l+1)} = d^{(k,l)} + \lambda L \beta^{(k,l+1)} - u^{(k,l+1)}. \quad (13)
\]

Note that the solution to (12) can be obtained directly using the soft thresholding operator

\[S(x, \delta) = \text{sgn}(x) \max\{0, |x| - \delta\},\]

while the solution to (11) can be obtained by applying a coordinate descent algorithm, which reduces to iteratively applying the soft thresholding operator for each component of \(\beta\) until convergence.

B List of ZCTAs in cancer clusters

We list only the ZCTAs in the four main clusters discussed in Sections 4.1 and 4.2. To facilitate comparisons, we separately list the ZCTAs that appear as part of each cluster under both models, and then separately list those that appear under only one of them.

B.1 Palm Beach

<table>
<thead>
<tr>
<th>ZCTAs that appear under both models</th>
<th>Appear only without covariate adjustment</th>
<th>Appear only with covariate adjustment</th>
</tr>
</thead>
<tbody>
<tr>
<td>33063 33067 33071 33403 33404</td>
<td></td>
<td>33498</td>
</tr>
<tr>
<td>33407 33408 33410 33411 33412</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33413 33426 33428 33434 33435</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33436 33437 33444 33445 33462</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33463 33467 33484 33496</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

B.2 Cape Coral

<table>
<thead>
<tr>
<th>ZCTAs that appear under both models</th>
<th>Appear only without covariate adjustment</th>
<th>Appear only with covariate adjustment</th>
</tr>
</thead>
<tbody>
<tr>
<td>33901 33903 33904 33905 33907</td>
<td></td>
<td>33440</td>
</tr>
<tr>
<td>33908 33909 33912 33914 33916</td>
<td></td>
<td>33471</td>
</tr>
<tr>
<td>33917 33919 33920 33922 33924</td>
<td></td>
<td>33930</td>
</tr>
<tr>
<td>33935 33936 33950 33955 33956</td>
<td></td>
<td>33931</td>
</tr>
<tr>
<td>33957 33971 33972 33990 33991</td>
<td></td>
<td>34134</td>
</tr>
<tr>
<td>33993</td>
<td>34142</td>
<td></td>
</tr>
</tbody>
</table>
B.3 Miami

<table>
<thead>
<tr>
<th>ZCTAs that appear under both models</th>
<th>Appear only without covariate adjustment</th>
<th>Appear only with covariate adjustment</th>
</tr>
</thead>
<tbody>
<tr>
<td>33004 33012 33013 33014 33015</td>
<td>33160</td>
<td></td>
</tr>
<tr>
<td>33016 33018 33020 33021 33023</td>
<td>33180</td>
<td></td>
</tr>
<tr>
<td>33025 33026 33027 33028 33029</td>
<td>33030</td>
<td></td>
</tr>
<tr>
<td>33055 33056 33109 33125 33126</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33128 33129 33130 33131 33132</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33134 33135 33136 33137 33139</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33143 33144 33145 33146 33149</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33155 33156 33157 33158 33165</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33166 33170 33172 33173 33174</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33175 33176 33177 33178 33165</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33183 33184 33185 33186 33187</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33189 33190 33193 33194 33196</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33322 33323 33325 33326 33327</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33328 33330 33331 33332 33351</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

B.4 North Florida

<table>
<thead>
<tr>
<th>ZCTAs that appear under both models</th>
<th>Appear only without covariate adjustment</th>
<th>Appear only with covariate adjustment</th>
</tr>
</thead>
<tbody>
<tr>
<td>32008 32009 32011 32024 32025</td>
<td>32621</td>
<td>32607</td>
</tr>
<tr>
<td>32026 32034 32038 32040 32043</td>
<td></td>
<td>32131</td>
</tr>
<tr>
<td>32044 32046 32054 32055 32058</td>
<td></td>
<td>32112</td>
</tr>
<tr>
<td>32060 32061 32062 32063 32064</td>
<td></td>
<td>32193</td>
</tr>
<tr>
<td>32066 32068 32071 32072 32073</td>
<td></td>
<td>32187</td>
</tr>
<tr>
<td>32083 32091 32092 32094 32097</td>
<td></td>
<td>32681</td>
</tr>
<tr>
<td>32134 32139 32140 32147 32148</td>
<td></td>
<td>32664</td>
</tr>
<tr>
<td>32177 32202 32204 32205 32207</td>
<td></td>
<td>32631</td>
</tr>
<tr>
<td>32210 32211 32212 32216 32217</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32218 32219 32220 32221 32223</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32225 32226 32234 32254 32256</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32257 32258 32259 32359 32606</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32608 32609 32615 32616 32618</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32619 32622 32625 32626 32628</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32640 32643 32653 32656 32658</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32666 32669 32680 32693 32694</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32697 34470 34471 34474 34475</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34476 34479 34481 34482 34488</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

References


