College of Public Health

Developing Cancer Epidemiology Maps to Detect Hot Spots for Small Geographic Areas

Jacob Oleson

March 25, 2025

CAMSA

Cancer Analytics & Maps for Small Areas

Areal Units ⓘ

ZCTA

Measure ⓘ

Age-adjusted Rate

Cancer ⓘ

Prostate

Outcome ⓘ

Incidence

Stratification ⓘ

Year Group

Year Group ⓘ

2016-2020

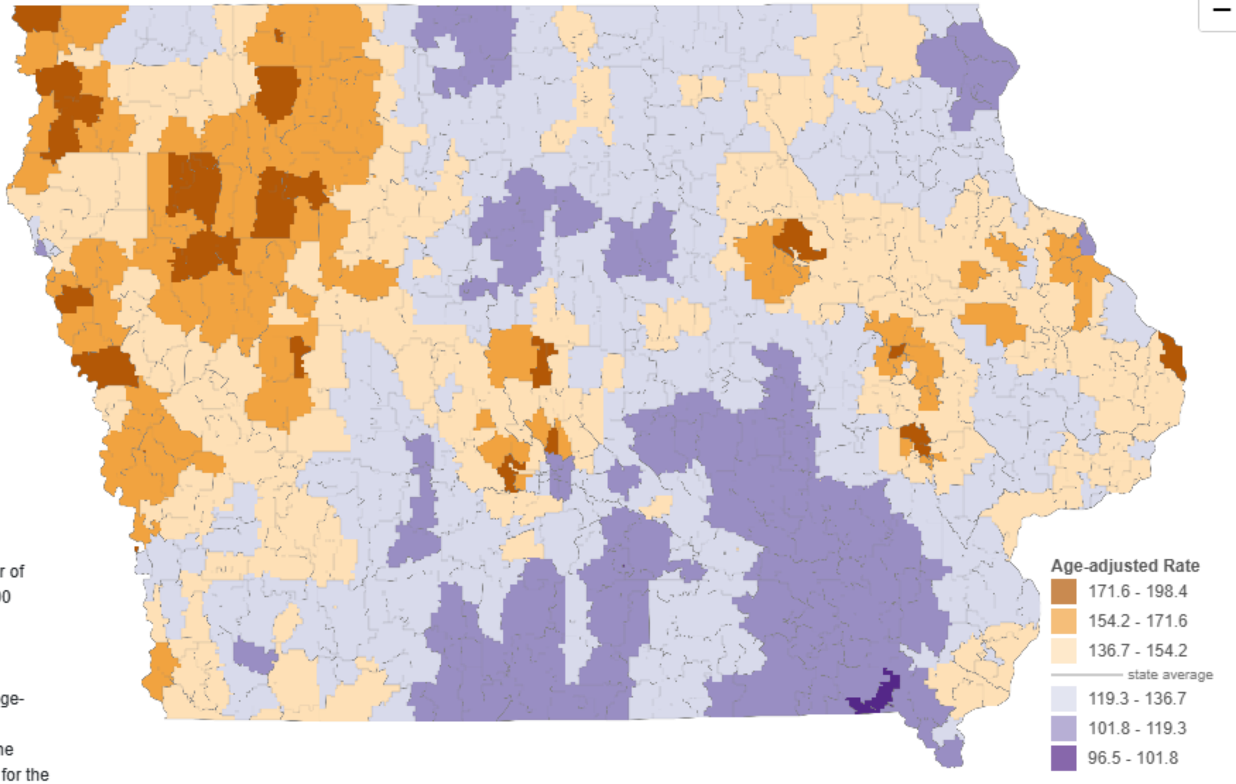
Filter by ⓘ

None

Update

Age-adjusted Prostate Cancer Incidence Rates (2016-2020)

Hover over an area unit to see details



This map shows the estimated number of new prostate cancer cases per 100,000 people in the 2016-2020 group, after adjusting for age.

For example, say a ZIP code has an age-adjusted rate of 58. This means, after adjusting for age, it is estimated that the number of new prostate cancer cases for the ZIP code is 58 per 100,000 people.

Search:

ZCTA	City	Age-adjusted Rate	AAR Standard Deviation	Risk Probability	Population Density (/km2)
50001	Ackworth	128.7824	18.9515	0.4833	5.4719
50002	Adair	130.2158	19.6702	0.5117	2.2832
50003	Adel	146.5837	15.7012	0.885	10.7639
50005	Albion	143.6043	25.1656	0.7143	5.6265
50006	Alden	120.8082	16.1544	0.307	2.7538
50007	Alleman	154.2391	22.7406	0.8793	17.9777
50008	Allerton	108.2772	18.9999	0.145	1.9501
50009	Altoona	146.9269	14.5636	0.9097	123.2708

Project Goals

- Small area estimates?
  - ZCTA, Census Tract, County
- Rural population?
  - Small counts
- Model estimates?
  - Useful results
  - Adjust for factors?

Where we started

- MS Preceptorship for Caitlin Ward

Applied Spatial Analysis and Policy
https://doi.org/10.1007/s12061-018-9276-4

CrossMark

Showcasing Cancer Incidence and Mortality in Rural ZCTAs Using Risk Probabilities via Spatio-Temporal Bayesian Disease Mapping

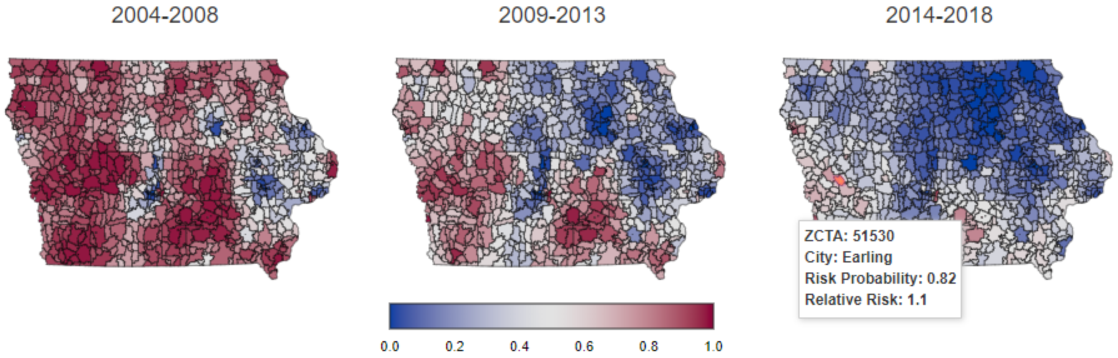
Caitlin Ward[1] · Jacob Oleson[1] · Katie Jones[2] · Mary Charlton[3]

- Create maps using spatial statistical methods that quantify cancer risk and are easy to understand
- Identify geographic disparities and potential areas to target with prevention programs
- Depict change over time 2004-2007, 2008-2011, 2012-2015

Search

About | Cancer Data | Research | Health Care Providers | Registrars | ICAN | Contact

Breast | Cervical | Colorectal | Liver | Lung | NHL | Prostate | Melanoma



Relative Risk Distribution by Time Period

Time Period	Minimum	25th Percentile	Median	Average	75th Percentile	Maximum	Standard Deviation
2004-2008	0.72	1.04	1.13	1.13	1.21	1.59	0.12
2009-2013	0.82	0.96	1.00	1.00	1.05	1.25	0.07
2014-2018	0.80	0.91	0.95	0.95	0.99	1.15	0.06
Total	0.72	0.94	1.00	1.03	1.09	1.59	0.11

Data

- Zip Code Tabulation Area (ZCTA), age, sex, year of diagnosis/death stage for every cancer in Iowa 2004-2015
  – Breast, cervical, colorectal, liver, lung, non-Hodgkin lymphoma, melanoma, prostate
  – 935 ZCTAs in Iowa
- Colorectal: 21,651 total cases
  – 2004-2007: 7,787 cases, 747 ZCTAs with < 10 cases
  – 2008-2011: 7,137 cases, 759 ZCTAs with < 10 cases
  – 2012-2015 6,727: cases, 766 ZCTAs with < 10 cases
- 2010 U.S. Census Bureau ZCTA level population in Iowa stratified by age and sex

Hierarchical Bayesian Poisson Model

- Assume ZCTA observed counts, O_{ij}, follow a Poisson distribution conditional on the SIR for each ZCTA and time-period

$$O_{ij} \sim Poisson(E_i * \theta_{ij})$$

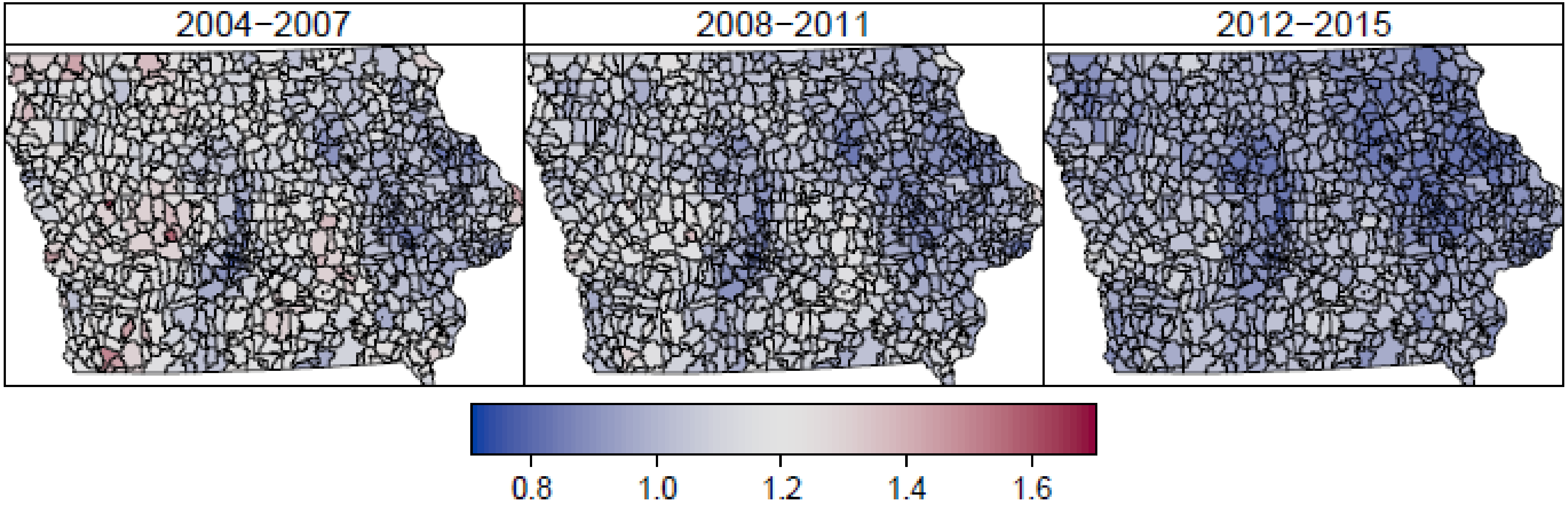
- A log-normal model was implemented for θ_{ij}
  - $\log(\theta_{ij}) = \beta_0 + Z_i + T_j + e_{ij}$ $DIC = 11140$
  - $\log(\theta_{ij}) = \beta_0 + u_i + v_i + T_j + e_{ij}$ $DIC = 11170$
  - $\log(\theta_{ij}) = \beta_0 + Z_i + T_j + \delta_i T_j$ $DIC = 11100$

Priors

- The spatial effects, Z_i, u_i, δ_i, are an intrinsic conditionally auto-regressive (ICAR)
  - Allows each ZCTA to borrow strength from its neighboring ZCTAs
- The correlation over time T_j is specified according to an autoregressive model (AR(1))

$$T_j = \rho_t T_{j-1} + e_{ij}$$

Colorectal Cancer Results Relative Risk

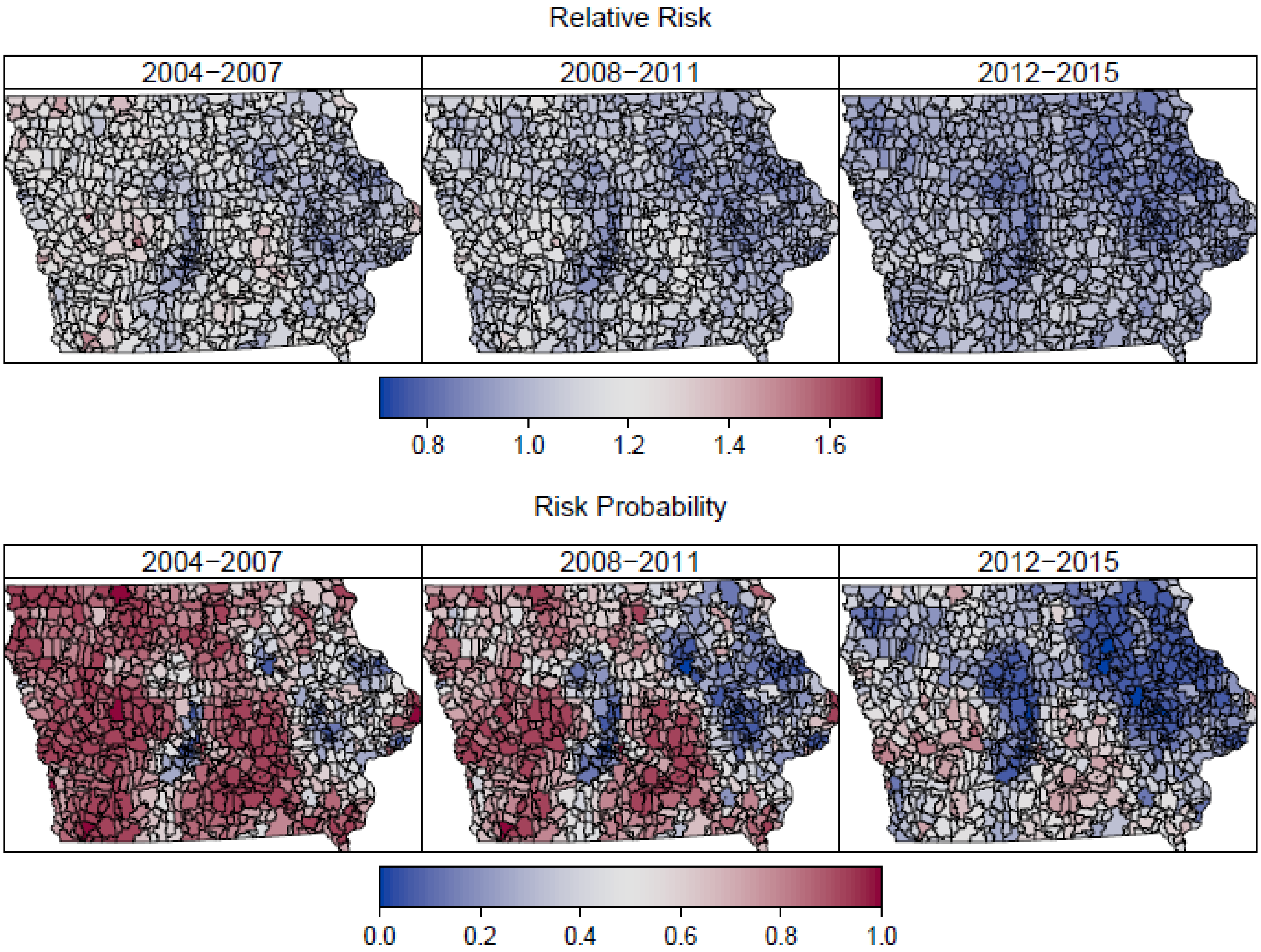


Risk Probability

- We were interested in computing a meaningful probabilistic quantity of risk
- The probability measure puts all maps for all cancers on the same scale (0.0 to 1.0)
- Bayesian exceedance probability computed as part of MCMC process

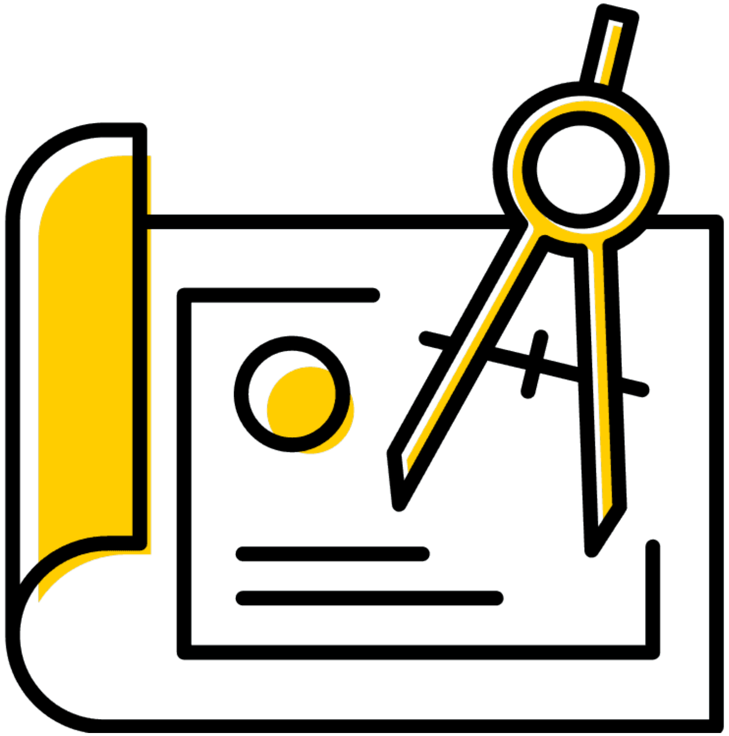
$$RP_{ij} = \frac{1}{K}\sum_{k=1}^{K} I(\theta_{ijk} > \gamma)$$

- We set our threshold to $\gamma = 1.0$


Relative Risk
2004−2007
2008−2011
2012−2015
0.8
1.0
1.2
1.4
1.6
Risk Probability
2004−2007
2008−2011
2012−2015
0.0
0.2
0.4
0.6
0.8
1.0

Goals of ITCR U01

- Provide age adjusted cancer rates per zip-code
- Identify hot spots and cold spots
- Account for demographics
  - Sex
  - Race
- Scalable
  - Produce for other states
  - Allow other states to upload their data



- The proportion of zero counts in the datasets are further inflated when the data are stratified by age group

Cancer	Proportion of zeros before age group stratification	Proportion of zeros after age group stratification
Liver	0.41	0.88
Colorectal	0.04	0.63

- We propose a Bayesian hierarchical hurdle model for estimating age-adjusted rates in disease mapping with excess zeros

RESEARCH ARTICLE

Statistics in Medicine WILEY

A Bayesian approach for estimating age-adjusted rates for low-prevalence diseases over space and time

Melissa Jay[1] | Jacob Oleson[1] | Mary Charlton[2] | Ali Arab[3]

Measures of risk in disease mapping

- Crude rate= $\frac{\#\ of\ deaths}{population\ size} * 100{,}000$
  - Often highly dependent on the underlying age distribution
- Age-adjusted rate= $\sum_{k=1}^{K} w_k * (crude\ rate\ for\ age\ group\ k)$
  - Direct standardization
  - Weighted average of age-group-specific crude rates
  - w_k's reflect the proportion of individuals in age group k in a selected standard population
  - Each age-group-specific rate is modeled separately and combined into an age-adjusted rate afterward

Bayesian Hierarchical Hurdle Model

- Stage 1: The probability of a non-zero count, π_{ijk}, is modeled using a Bernoulli regression model
- Stage 2: Positive counts are modeled using a zero-truncated Poisson regression model with parameter θ_{ijk}
- $P\left(Y_{ijk} = y_{ijk} \middle| \pi_{ijk}, \theta_{ijk}\right) = \begin{cases} 1 - \pi_{ijk} & y_{ijk} = 0 \\ \pi_{ijk} * \dfrac{\theta_{ijk}^{y_{ijk}} \exp\{-\theta_{ijk}\}}{y_{ijk}!(1-\exp\{-\theta_{ijk}\})} & y_{ijk} > 0 \end{cases}$

Linear Predictors in Regression Models

- Stage 1: Bernoulli Regression Model

$$g(\pi_{ijk}) = \boldsymbol{\alpha}_1 \boldsymbol{x}_k + \beta_{pop} \log(n_{ijk}) + \boldsymbol{\alpha}_2 \log(n_{ijk}) * \boldsymbol{x}_k + \gamma_{1i} + \delta_{1j}$$

- $g(\pi_{ijk})$ is the complementary log log link $(\log[-\log(1-\pi)])$
- $\boldsymbol{x}_k$ contains the age information (dummy variables) using 6 age groups
- γ_{1i} is a spatial random effect (ICAR)
- δ_{1j} is a temporal random effect
- Log of the population size is included as a covariate in stage 1 since there is not a natural way to incorporate a population offset

Linear Predictors in Regression Models

- Stage 1: Bernoulli Regression Model

$$g(\pi_{ijk}) = \boldsymbol{\alpha}_1 \boldsymbol{x}_k + \beta_{pop} \log(n_{ijk}) + \boldsymbol{\alpha}_2 \log(n_{ijk}) * \boldsymbol{x}_k + \gamma_{1i} + \delta_{1j}$$

- Stage 2: Zero-truncated Poisson Regression Model

$$\log(\theta_{ijk}) = \log(n_{ijk}) + \boldsymbol{x}_k^T \boldsymbol{\beta} + \gamma_{2i} + \delta_{2j} + \epsilon_{ij}$$

- $\boldsymbol{x}_k$ contains the age information (dummy variables) using 6 age groups
- γ_{1i} and γ_{2i} are spatial random effects (ICAR)
- δ_{1j} and δ_{2j} are temporal random effects
- ϵ_{ij} accounts for uncorrelated heterogeneity

Additional Priors

- Age group coefficients
  - α and β have diffuse independent normal priors
- Random effects
  - $\gamma_1 \sim ICAR(\tau_{y1})$
  - $\gamma_2 \sim ICAR(\tau_{y2})$
  - $\delta_1 \sim N(0, \sigma_{d1}^2)$
  - $\delta_2 \sim N(0, \sigma_{d2}^2)$
- Hyperparameters
  - Correlation parameters have Uniform(-1,+1) priors
  - Standard deviation parameters have Half-Cauchy(10) priors

Age-adjusted rates

- The age-group-specific rate for region i during year j is calculated by dividing $E(Y_{ijk})$ by n_{ijk} and then multiplying by 100,000 individuals

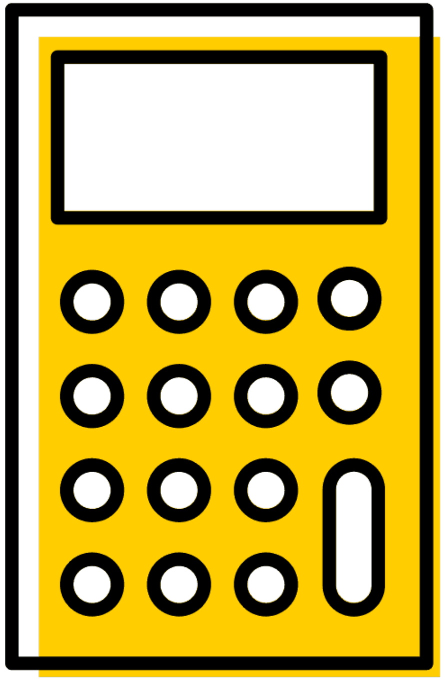
$$R_{ijk} = \left(\frac{\frac{\pi_{ijk} * \theta_{ijk}}{1 - \exp\{-\theta_{ijk}\}}}{n_{ijk}} \right) * 100{,}000$$

- Thus, the age-adjusted rate for each region and year is

$$R_{ij} = \sum_{k=1}^{K} w_k * R_{ijk}$$

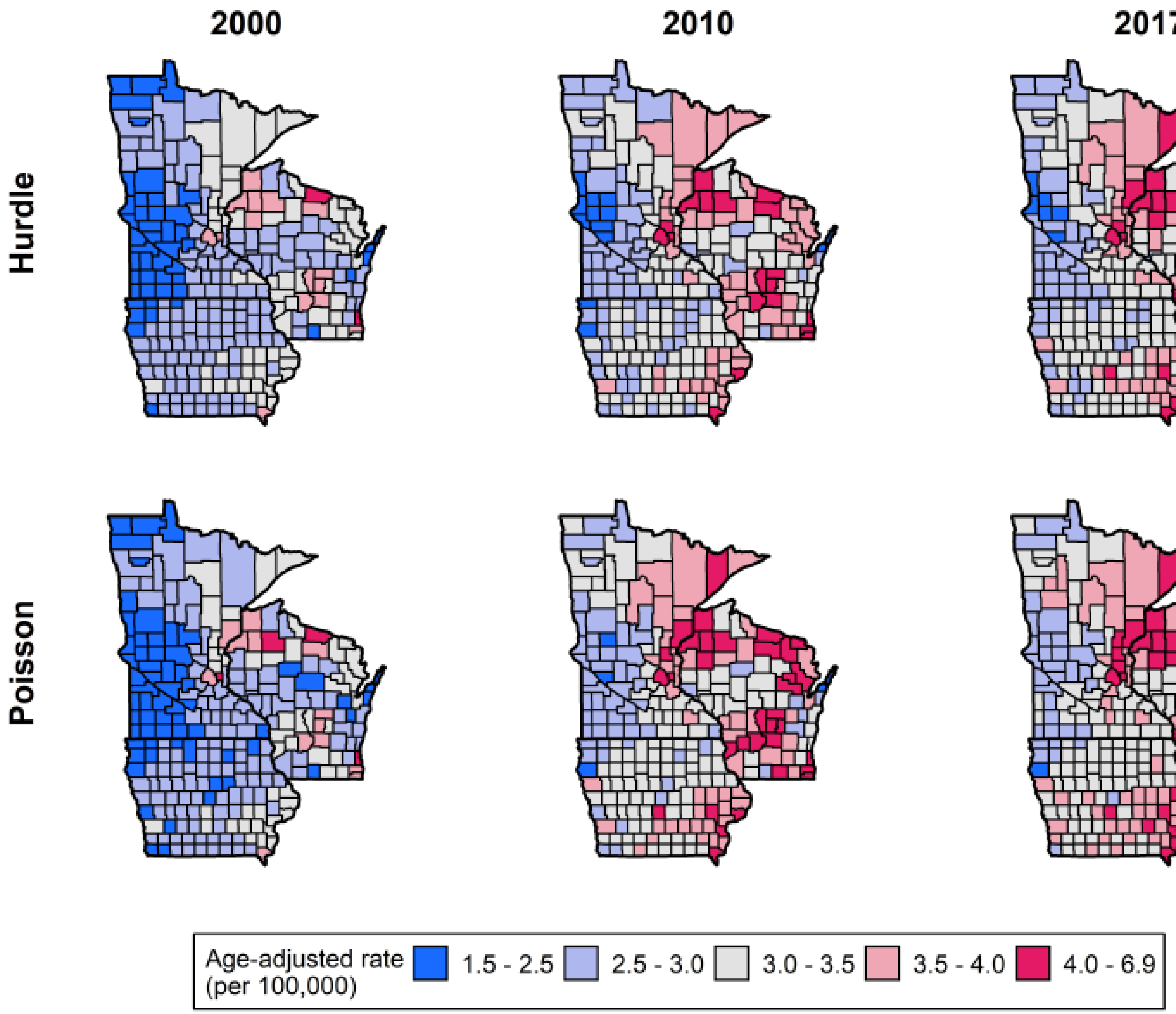
- We obtain 1,000 posterior samples of π_{ijk} and θ_{ijk} to compute the posterior mean and variance of each age-adjusted rate

Age-adjusted rates



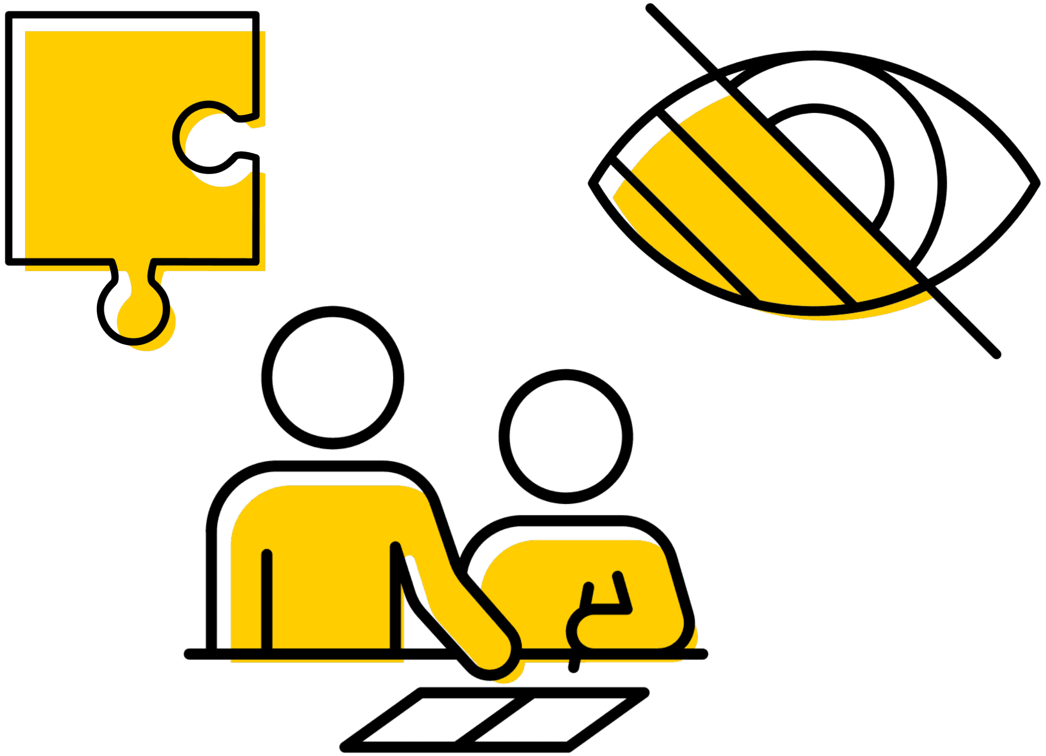
- Bayesian exceedance probabilities computed the same way as before for hot spots
- General code in R
  - We have code in OpenBUGS, INLA, and NIMBLE
  - Compute crude rates, age-adjusted rates, cancer risk, specify cancer
- Mapping performed in R, ArcGIS, Leaflet

Liver cancer



Additional Goals Since Getting Started: Missing puzzle pieces

- More useful for cancer patients, survivors, friends, and family
- More Color Options
- Better Descriptions
- Better Graphics
- Subset by Sex, Race



Current Tool Capabilities: version 2.0

- Interactive web-based application
- Eight cancers:
  - Colorectal, breast, cervical, liver, lung, melanoma, NHL, prostate
- Geographic units:
  - ZIP code tabulation areas (ZCTA), County
- Stratification options:
  - Sex, year group, race/ethnicity
- Different measures:
  - Age-adjusted rate, risk probability, population estimate
- Highlight clusters/hot spots

CAMSA

Cancer Analytics & Maps for Small Areas

Areal Units
ZCTA

Measure
Age-adjusted Rate

Cancer
Prostate

Outcome
Incidence

Stratification
Year Group

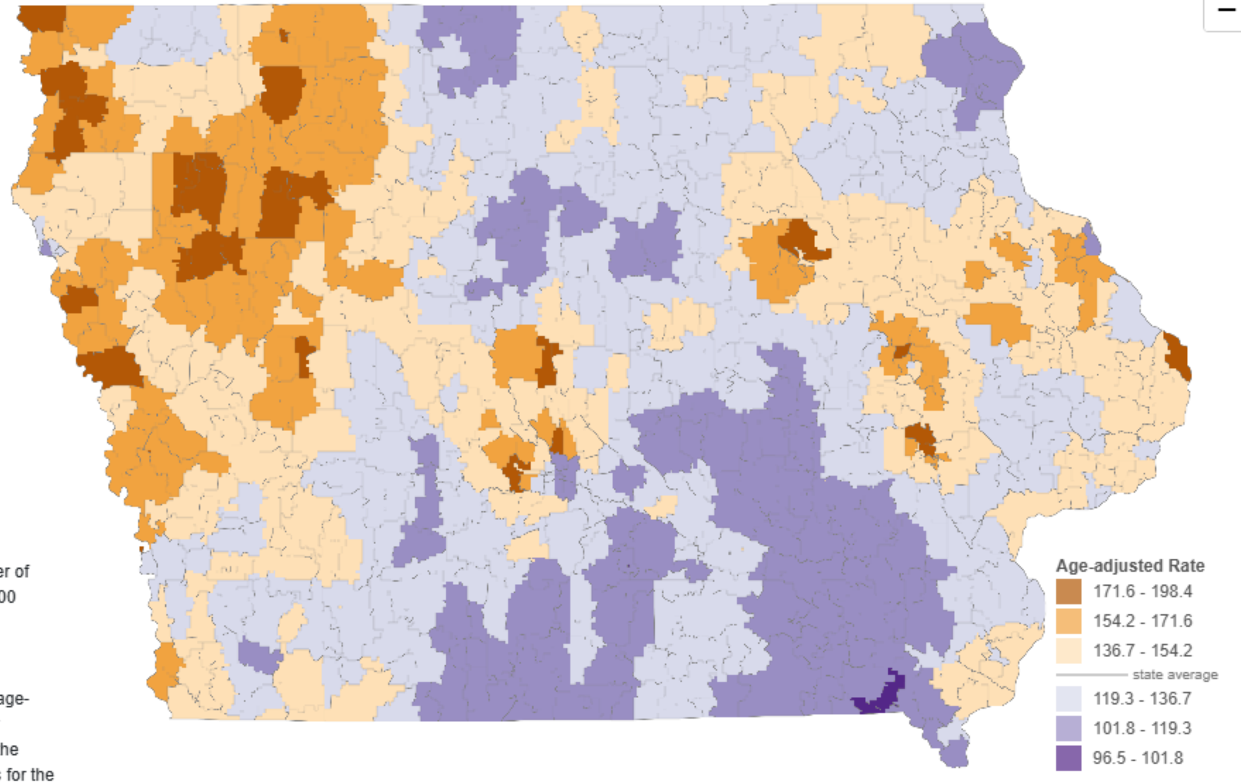
Year Group
2016-2020

Filter by
None

Update

Age-adjusted Prostate Cancer Incidence Rates (2016-2020)

Hover over an area unit to see details



This map shows the estimated number of new prostate cancer cases per 100,000 people in the 2016-2020 group, after adjusting for age.

For example, say a ZIP code has an age-adjusted rate of 58. This means, after adjusting for age, it is estimated that the number of new prostate cancer cases for the ZIP code is 58 per 100,000 people.

Search:

ZCTA	City	Age-adjusted Rate	AAR Standard Deviation	Risk Probability	Population Density (/km2)
50001	Ackworth	128.7824	18.9515	0.4833	5.4719
50002	Adair	130.2158	19.6702	0.5117	2.2832
50003	Adel	146.5837	15.7012	0.885	10.7639
50005	Albion	143.6043	25.1656	0.7143	5.6265
50006	Alden	120.8082	16.1544	0.307	2.7538
50007	Alleman	154.2391	22.7406	0.8793	17.9777
50008	Allerton	108.2772	18.9999	0.145	1.9501
50009	Altoona	146.9269	14.5636	0.9097	123.2708

Age Adjusted Rates

Areal Units ⓘ
ZCTA

Measure ⓘ
Age-adjusted Rate

Cancer ⓘ
Colorectal

Outcome ⓘ
Incidence

Stratification ⓘ
Year Group

Year Group ⓘ
2006-2010

Filter by ⓘ
None

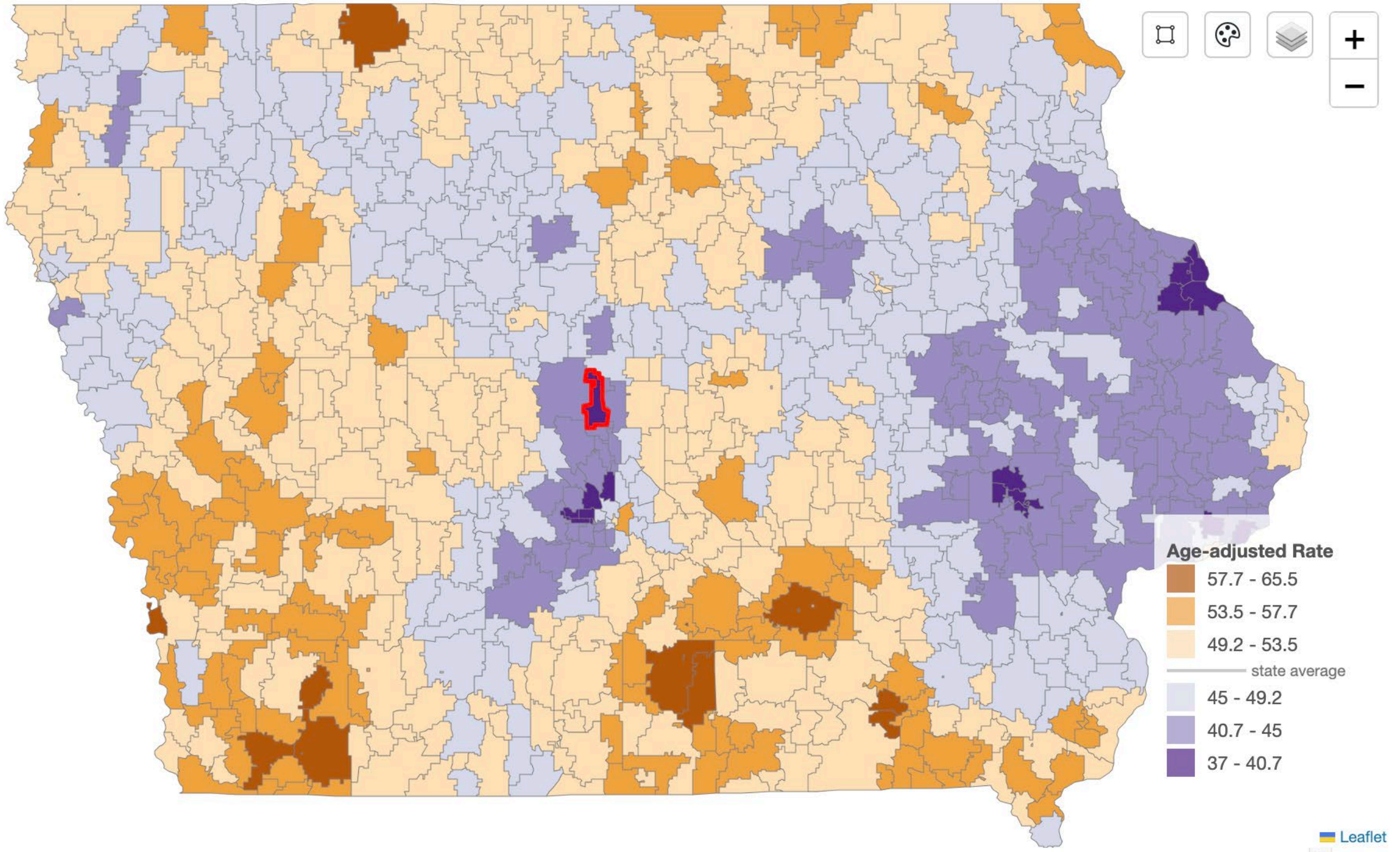
Update

ZCTA: 50014
City: Ames
Age-adjusted Rate: 39.5684
AAR Standard Deviation: 3.9128
Risk Probability: 0.0853
Population Density (/km2): 193.4961

This map shows the estimated number of new colorectal cancer cases per 100,000 people in the 2006-2010 group, after adjusting for age.

For example, say a ZIP code has an age-adjusted rate of 58. This means, after adjusting for age, it is estimated that the number of new colorectal cancer cases for the ZIP code is 58 per 100,000 people.

Age-adjusted Colorectal Cancer Incidence Rates (2006-2010)



Risk Probability

Areal Units ⓘ

ZCTA

Measure ⓘ

Risk Probability

Cancer ⓘ

Female Breast

Outcome ⓘ

Incidence

Stratification ⓘ

Year Group

Year Group

2016-2020

Filter by ⓘ

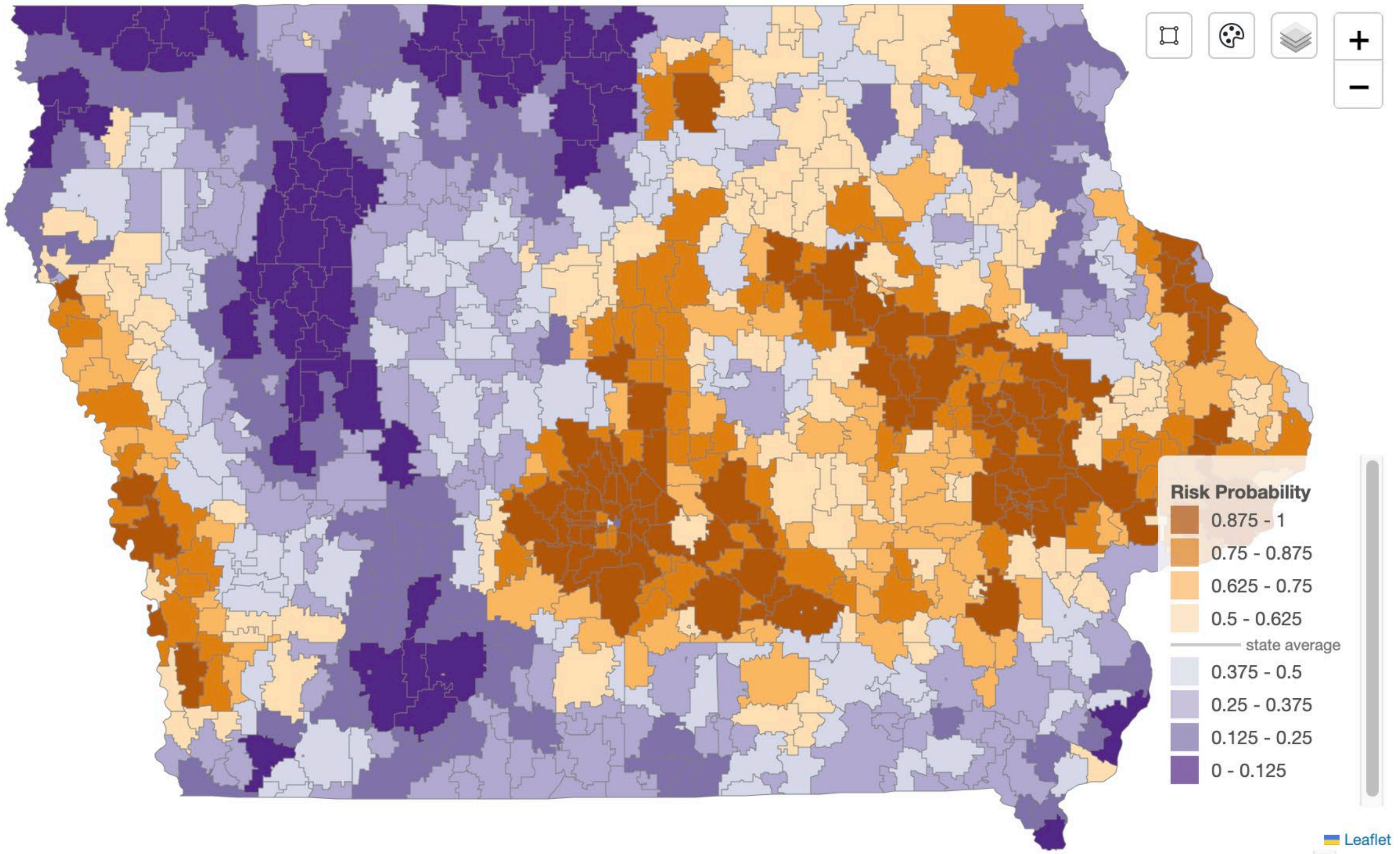
None

Update

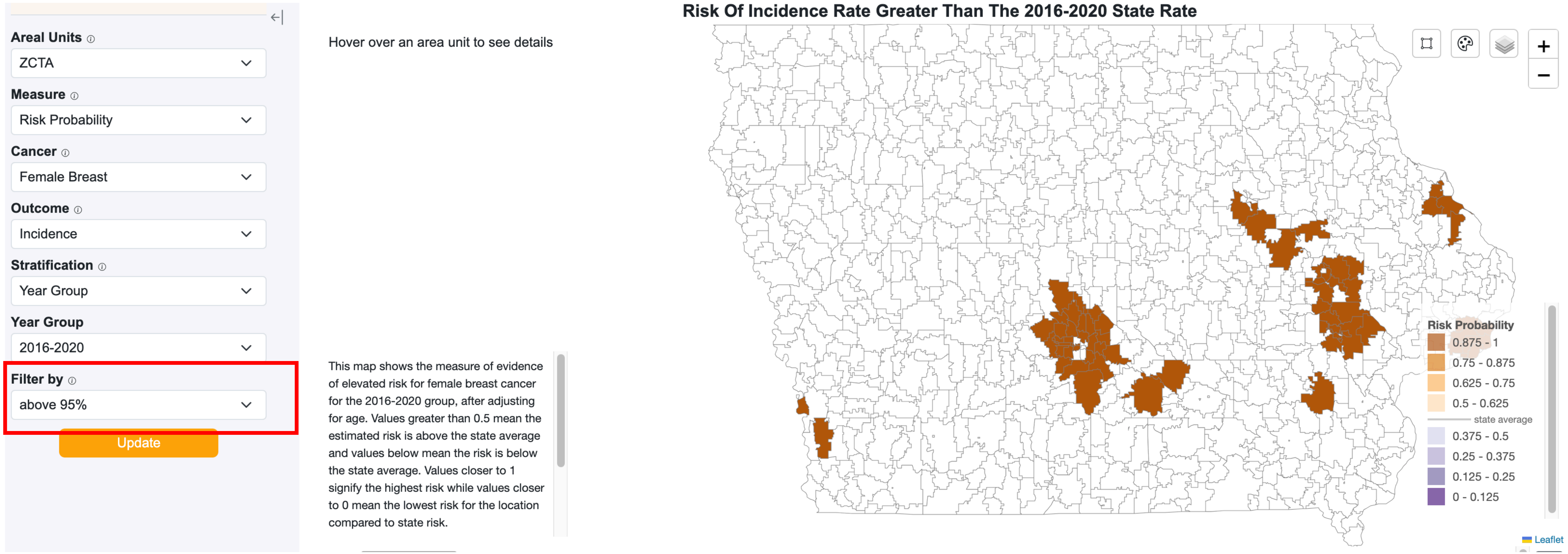
Hover over an area unit to see details

This map shows the measure of evidence of elevated risk for female breast cancer for the 2016-2020 group, after adjusting for age. Values greater than 0.5 mean the estimated risk is above the state average and values below mean the risk is below the state average. Values closer to 1 signify the highest risk while values closer to 0 mean the lowest risk for the location compared to state risk.

Risk Of Incidence Rate Greater Than The 2016-2020 State Rate



Hot Spots



Areal Units
ZCTA
Measure
Risk Probability
Cancer
Female Breast
Outcome
Incidence
Stratification
Year Group
Year Group
2016-2020
Filter by
above 95%
Update
Hover over an area unit to see details
This map shows the measure of evidence of elevated risk for female breast cancer for the 2016-2020 group, after adjusting for age. Values greater than 0.5 mean the estimated risk is above the state average and values below mean the risk is below the state average. Values closer to 1 signify the highest risk while values closer to 0 mean the lowest risk for the location compared to state risk.
Risk Of Incidence Rate Greater Than The 2016-2020 State Rate
Risk Probability
0.875 - 1
0.75 - 0.875
0.625 - 0.75
0.5 - 0.625
state average
0.375 - 0.5
0.25 - 0.375
0.125 - 0.25
0 - 0.125
Leaflet

Layering features

Areal Units

ZCTA

Measure

Risk Probability

Cancer

Female Breast

Outcome

Incidence

Stratification

Year Group

Year Group

2016-2020

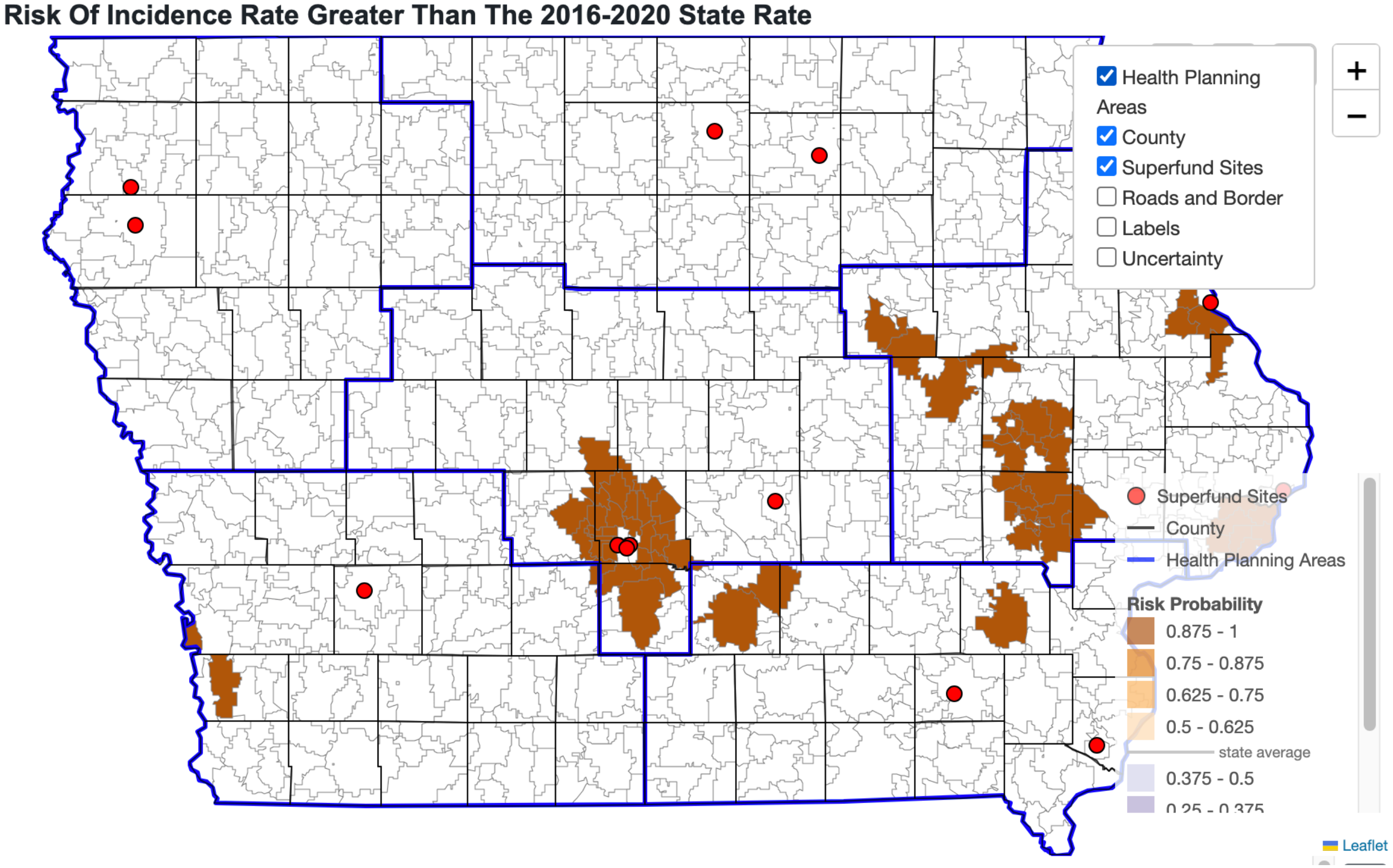
Filter by

above 95%

Update

Hover over an area unit to see details

This map shows the measure of evidence of elevated risk for female breast cancer for the 2016-2020 group, after adjusting for age. Values greater than 0.5 mean the estimated risk is above the state average and values below mean the risk is below the state average. Values closer to 1 signify the highest risk while values closer to 0 mean the lowest risk for the location compared to state risk.



Acknowledgements

Research Team

Jake Oleson (PI)

Sarah Nash

Grant Brown

Carly Mahoney

Jacob Clark

Ali Zorn

Mary Charlton

Erin Wissler Gerdes

Caglar Koylu

Jinyi Cai

Brittany McKelvey

Funding and Data

This study is funded by the National Institutes of Health/National Cancer Institute: NCI 1 U01 CA258400-01A1

Data are from the Iowa Cancer Registry: NIH NCI Contract HHSN261201800012I HHSN2610001 and the New Mexico Tumor Registry: NIH NCI contract HHSN261201300010I, Task Order HHSN26100005

IOWA

IOWA

College of Public Health

Questions?

Jacob Oleson
Department of Biostatistics

https://shri.public-health.uiowa.edu/cancer-data/cancer-maps/iowa-cancer-maps/

319-384-1595
Jacob-Oleson@uiowa.edu

→ uiowa.edu