Modeling age-adjusted rates from spatio-temporal data sets with excess zero counts

Melissa Jay, MS
University of Iowa, Department of Biostatistics

Joint work with: Jacob Oleson, PhD, Mary Charlton, PhD, and Ali Arab, PhD

Outline

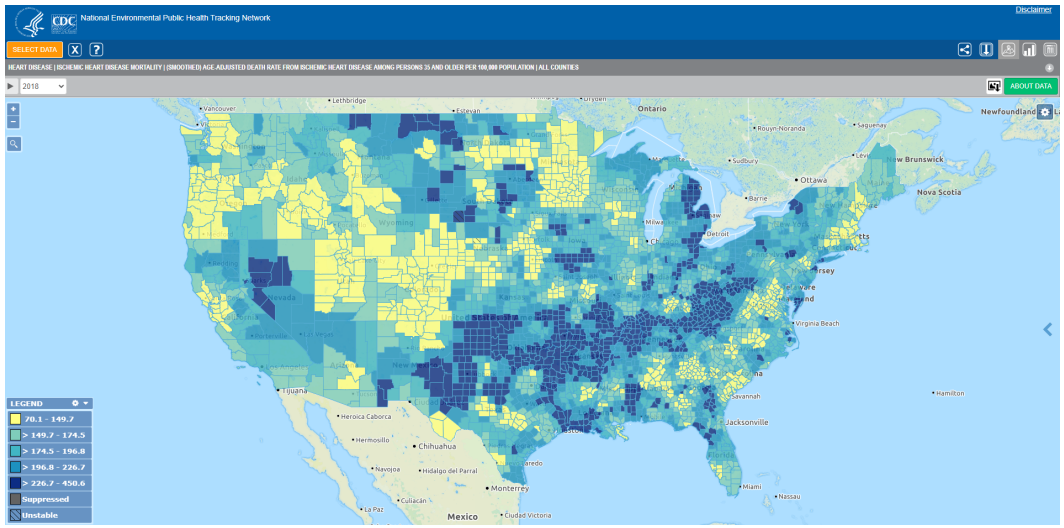
- Disease mapping
- Proposed hurdle model for age-adjusted rates
  - Model specification
  - Model implementation
  - Estimation of age-adjusted rates
- Application to county-level cancer mortality
- Simulation study
- Conclusions

Small area estimation

- Small area estimation (SAE) methods aim to estimate area-specific parameters for populations in small geographic regions
  - To do this, they might incorporate information from nearby regions or large surveys to produce more stable estimates for the small areas
- Government statistical agencies use SAE techniques to obtain county-level estimates of measures such as median household income, smoking rates, and literacy rates
- In this presentation, I will specifically focus on the application of a new SAE model for disease mapping, but the SAE models presented can be applied across disciplines

Disease mapping

- Disease mapping is an important tool in spatial epidemiology used to assess patterns in disease incidence or mortality across space and time
- By quantifying disease risk at the county level, health departments can efficiently allocate resources and promote prevention efforts in the communities that need them most



Estimates of age-adjusted mortality rates from ischemic heart disease in 2018. Accessed using the CDC's National Environmental Public Health Tracking Network data visualization tool.

Measures of risk in disease mapping

- **Crude rate** $= \frac{\#\text{ of deaths}}{\text{population size}} * 100,000$
  - Often highly dependent on the underlying age distribution
- **Age-adjusted rate** $= \sum_{k=1}^{K} w_k * (\text{crude rate for age group k})$
  - 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 calculated separately and combined into an age-adjusted rate afterward
- **Standardized mortality ratio (SMR)** $= \frac{\#\text{ of observed deaths}}{\#\text{ of expected deaths}}$
  - Indirect standardization
  - Expected deaths for a county and year are calculated upfront, so there is only one data point per county and year

Why should we model age-adjusted rates?

- For low-prevalence diseases and in counties with small population sizes, raw mortality rates are subject to considerable variability
  - Mapping the raw mortality rates directly might lead researchers to identify spurious patterns that are present due to variability
- For example, say a county has a population size of 1000 people
  - Rate if there were 0 deaths: (0/1000)*100,000 = 0
  - Rate if there was 1 death: (1/1000)*100,000 = 100
  - Rate if there were 2 deaths: (2/1000)*100,000 = 200
- Disease risk is often spatially and temporally correlated
  - We can borrow strength from nearby counties and years within a model to leverage this spatial and temporal correlation

Why should we model age-adjusted rates? (continued)

- In practice, age-adjusted rates are often calculated upfront then treated as a continuous outcome in a linear regression model. Some downsides of this approach include:
  - It ignores the variability in the rate calculation
  - It assigns a sample size of one to each areal unit
  - It is not possible to incorporate individual-level data into the analysis
- Bayesian hierarchical modeling approaches can:
  - Appropriately account for the variability in the rate calculations
  - Leverage the spatial and temporal dependencies in the dataset
  - Incorporate individual-level data or regional covariates into the estimation, if desired

Bayesian hierarchical Poisson regression model

- The Bayesian hierarchical Poisson regression model with spatial and temporal random effects is frequently used in disease mapping settings

$$\begin{aligned} Y_{i,j,k} &\sim \text{Poisson}(\theta_{i,j,k}) \\ \log(\theta_{i,j,k}) &= \log(n_{i,j,k}) + \boldsymbol{x_k^T}\boldsymbol{\beta} + \gamma_i + \delta_j + \epsilon_{i,j} \end{aligned}$$

- $Y_{i,j,k}$ is the number of deaths in county i during year j for age group k
- $n_{i,j,k}$ is the corresponding population size
- $\boldsymbol{x_k}$ is a $K \times 1$ vector of age group indicators corresponding to age group k
- γ_i and δ_j are spatial and temporal random effects
- $\epsilon_{i,j}$ accounts for overdispersion
- In addition to the likelihood, a set of prior distributions must be specified for the parameters, random effects, and hyperparameters

Age-adjusted rates

Age-adjusted rates can be estimated from the Poisson model by:

1. Drawing posterior samples of each $\theta_{i,j,k}$
2. Computing each age-group-specific rate for county i during year j as:

$$R_{i,j,k} = \frac{\theta_{i,j,k}}{n_{i,j,k}} * 100,000$$

3. Combining the age-group-specific rates into an age-adjusted rate using standard population weights:

$$R_{i,j} = \sum_{k=1}^{K} w_k * R_{i,j,k}$$

4. Computing the posterior mean of each $R_{i,j}$ to obtain a rate estimate for each county and year

Motivation for the proposed hurdle model

- This work is motivated by the need for reliable estimates of age-adjusted cancer mortality rates in the Midwest
- Since the Midwest has a particularly large proportion of rural counties, cancer mortality data sets include many zero counts
- The proportion of zero counts in the data sets 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

- The aforementioned Poisson model does not account for excess zeros
- We propose a **Bayesian hierarchical hurdle model** for estimating age-adjusted rates in disease mapping settings with excess zeros

Likelihood

- To model age-adjusted rates, we specify a hurdle model for the likelihood
- The hurdle model accounts for excess zeros using a two-stage approach
  - **Stage 1:** The probability of a non-zero count, $\pi_{i,j,k}$, is modeled using a Bernoulli regression model
  - **Stage 2:** Positive counts are modeled using a zero-truncated Poisson regression model with parameter $\theta_{i,j,k}$

$$P(Y_{i,j,k} = y_{i,j,k} \mid \pi_{i,j,k}, \theta_{i,j,k}) = \begin{cases} 1 - \pi_{i,j,k}, & y_{i,j,k} = 0 \\ \pi_{i,j,k} * \frac{\theta_{i,j,k}^{y_{i,j,k}} \exp\{-\theta_{i,j,k}\}}{y_{i,j,k}!(1-\exp\{-\theta_{i,j,k}\})}, & y_{i,j,k} > 0 \end{cases}$$

where i denotes the county, j denotes the year, and k denotes the age group

Stage 1: Bernoulli regression model

$$g(\pi_{i,j,k}) = \boldsymbol{x_k^T \alpha_1} + \log(n_{i,j,k}) * \boldsymbol{x_k^T \alpha_2} + \gamma_{1,i} + \delta_{1,j}$$

- The function g is left unspecified, since the complementary log-log or logit link work well, depending on the application
- $\boldsymbol{x_k}$ is a $K \times 1$ vector of age group indicators corresponding to age group k
- $\gamma_{1,i}$ is a spatial random effect for county i
- $\delta_{1,j}$ is a temporal random effect for year j
- Log of the population size is included as a covariate, since there is not a natural way to incorporate a population offset

Stage 2: Zero-truncated Poisson regression model

$$\log(\theta_{i,j,k}) = \log(n_{i,j,k}) + \boldsymbol{x_k^T}\boldsymbol{\beta} + \gamma_{2,i} + \delta_{2,j} + \epsilon_{i,j}$$

- $\gamma_{2,i}$ is a spatial random effect
- $\delta_{2,j}$ is a temporal random effect
- $\epsilon_{i,j}$ accounts for uncorrelated heterogeneity

Priors

Age group coefficients:

- $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ have diffuse independent normal priors

Random effects:

- $\boldsymbol{\gamma_1} \mid \tau_{\gamma_1} \sim ICAR(\tau_{\gamma_1})$
- $\boldsymbol{\gamma_2} \mid \tau_{\gamma_2} \sim ICAR(\tau_{\gamma_2})$
- $\boldsymbol{\delta_1} \mid \rho_1, \tau_{\delta_1} \sim AR(1)$ with correlation ρ_1 and precision τ_{δ_1}
- $\boldsymbol{\delta_2} \mid \rho_2, \tau_{\delta_2} \sim AR(1)$ with correlation ρ_2 and precision τ_{δ_2}
- $\boldsymbol{\epsilon} \mid \tau_\epsilon \sim \text{Normal}(\mathbf{0}, \tau_\epsilon * \boldsymbol{I})$

Hyperparameters:

- Correlation parameters ρ_1 and ρ_2 have Uniform(-1,1) priors
- Standard deviation parameters have Half-Cauchy(10) priors

Spatial process

- To allow for spatial smoothing, we apply independent intrinsic conditional autoregressive (ICAR) priors to the spatial random effects in each stage of the hurdle model
- The ICAR prior treats the spatial random effects for a county and its neighbors as correlated

$$\gamma_{1,i} \mid \boldsymbol{\gamma_{1,-i}} \sim \text{Normal}\left(\frac{\sum_{i\sim h}\gamma_{1,h}}{m_i}, m_i\tau_{\gamma_1}\right)$$

where m_i = number of counties adjacent to county i and $i \sim h$ represents adjacency between counties i and h

- To ensure identifiability, $\sum_{i=1}^{I}\gamma_{1,i} = 0$

Temporal process

- To allow for temporal smoothing, we apply independent autoregressive(1) priors to the temporal random effects in each stage of the hurdle model
- The AR(1) prior treats the temporal random effects for a given year and the adjacent years as correlated

$$\delta_{1,0} \sim \text{Normal}(0, \tau_{\delta_1}(1-\rho_1^2))$$

$$\delta_{1,j} = \rho_1 * \delta_{1,j-1} + \kappa_{1,j} \text{ for } j = 2, \ldots, J$$

where $\kappa_{1,j} \sim \text{Normal}(0, \tau_{\delta_1})$

Age-adjusted rates

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

$$R_{i,j,k} = \left(\frac{\pi_{i,j,k} * \theta_{i,j,k}}{1 - \exp\{-\theta_{i,j,k}\}} \Big/ n_{i,j,k} \right) * 100,000$$

- Thus, the age-adjusted rate for each county and year is computed as:

$$R_{i,j} = \sum_{k=1}^{K} w_k * R_{i,j,k}$$

- We obtain 1,000 posterior samples of each $\pi_{i,j,k}$ and $\theta_{i,j,k}$ to compute the posterior mean and variance of each age-adjusted rate

Model implementation

- This modeling approach can quickly become computationally expensive with the addition of counties and years
- We use the INLA package in R to run these models
  - Approximate Bayesian inference greatly reduces model run time compared to MCMC methods

Software	**Runtime (Iowa data)**
INLA	$\sim$ 3.5 minutes
OpenBUGS	$\sim$ 12 hours

- Since the posterior distribution can be factored, each stage can be run in parallel to further reduce run time

Hurdle model in INLA - stage 1

```
y1 <- ifelse(data$deaths > 0, 1, 0)

fit1 <- inla(y1 ~ 0 + mu + age1 + age2 + age3 + age4 +
               logpop + age1*logpop + age2*logpop + age3*logpop + age4*logpop +
               f(county_id, model = "besag", graph = adj,
                 hyper = list(prec = list(prior = HC.prior))) +
               f(year_id, model = "ar1",
                 hyper = list(theta1 = list(prior = HC.prior),
                              theta2 = list(prior = "betacorrelation",
                                            param = c(1,1)))),
             data = dat1, family = "binomial",
             control.family = list(link = "cloglog"),
             control.compute = list(dic = TRUE, waic = TRUE, config = TRUE))

## Function to compute pi from fit1 predictor
fun1 <- function(){
  1-exp(-exp(Predictor))
}
```

Hurdle model in INLA - stage 2

Type 0

The (type 0) likelihood is defined as

$$\text{Prob}(y \mid \ldots) = p \times 1_{[y=0]} + (1-p) \times \text{Poisson}(y \mid y > 0)$$

where p is a hyperparameter where

$$p = \frac{\exp(\theta)}{1+\exp(\theta)}$$

and θ is the internal representation of p; meaning that the initial value and prior is given for θ. This is model is called `zeroinflatedpoisson0` (and `zeroinflatedbinomial0`).

- The zero-truncated Poisson model is not implemented in INLA
- To run stage 2, we use the `zeroinflatedpoisson0` model in INLA and set $\theta = -20$ so that $p \approx 0$

Hurdle model in INLA - stage 2 (continued)

```
y2 <- ifelse(data$deaths == 0, NA, data$deaths)

fit2 <- inla(y2 ~ 0 + offset(logpop) + mu + age1 + age2 + age3 + age4 +
               f(county_id, model = "besag", graph = adj,
                 hyper = list(prec = list(prior = HC.prior))) +
               f(year_id, model = "ar1",
                 hyper = list(theta1 = list(prior = HC.prior),
                              theta2 = list(prior = "betacorrelation",
                                            param = c(1,1)))) +
               f(county_year_id, model = "iid",
                 hyper = list(prec = list(prior = HC.prior))),
             data = dat2, family = c("zeroinflated.poisson0"),
             control.family = list(list(
               hyper = list(prob = list(initial = -20, fixed = TRUE)))),
             control.compute = list(dic = TRUE, waic = TRUE, config = TRUE))

## Function to compute theta from fit2 predictor
fun2 <- function(){
  exp(Predictor)
}
```

Hurdle model in INLA - estimating age-adjusted rates

```
## Draw 1000 samples of pi from fit1
nSamp <- 1000
set.seed(211)
samp1 <- inla.posterior.sample(nSamp, fit1, seed = 211)
pi_samp <- inla.posterior.sample.eval(fun1, samp1)

## Draw 1000 samples of theta from fit2
set.seed(211)
samp2 <- inla.posterior.sample(nSamp, fit2, seed = 211)
theta_samp <- inla.posterior.sample.eval(fun2, samp2)

## Compute rates
rates <- (pi_samp /(1-exp(-theta_samp)))*theta_samp*100000/data$population

## Obtain mean rate for each county, year, and age group
rates <- rowMeans(rates)
rates <- matrix(rates, ncol = 5, byrow = T)

## Take weighted average using standard population weights
rates <- c(rates %*% weights)
```

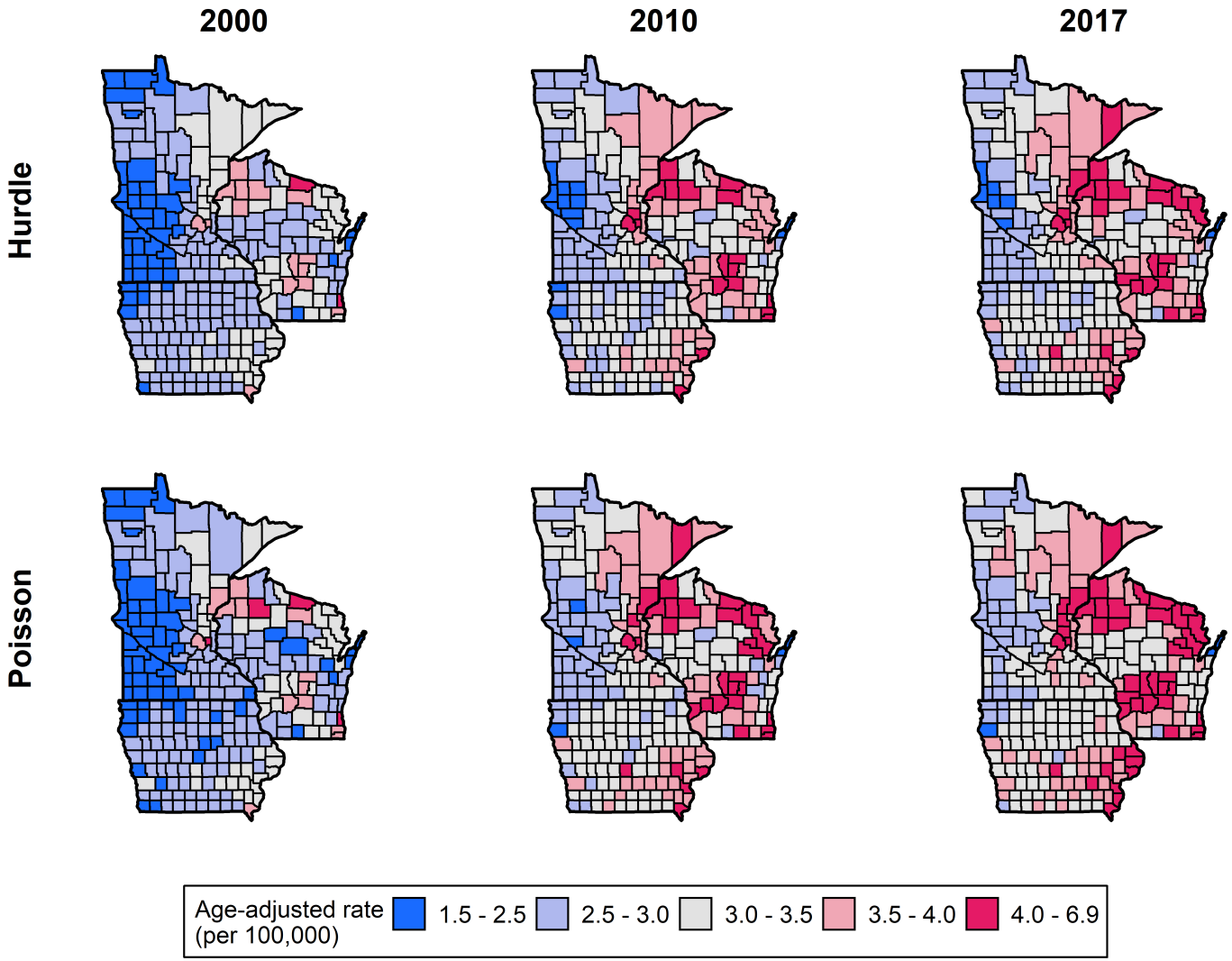
Application to county-level cancer mortality

- Annual cancer death counts were derived from the National Center for Health Statistics Vital Statistics data files and age-adjusted using the 2010 U.S. standard population
  - **Counties:** All counties in Iowa, Minnesota, and Wisconsin ($I = 258$)
  - **Years:** 2000 - 2017 ($J = 18$)
  - **Age groups:** $<$ 40, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80-84, and 85+ ($K = 11$)

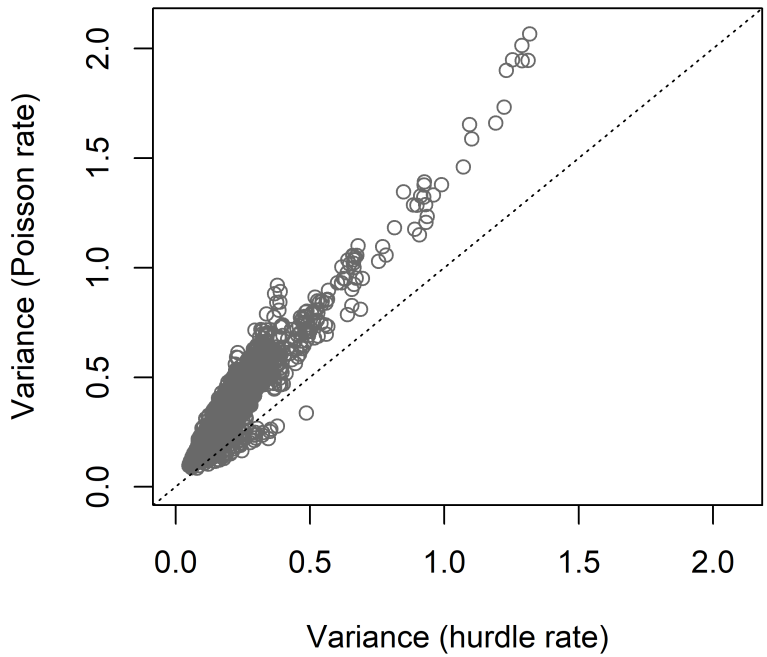
Cancer	**Proportion of zeroes**	**Mean of non-zero counts**	**Median of non-zero counts**	**Pattern over time**
Liver	0.88	1.5	1.0	Increasing
Colorectal	0.63	2.3	1.0	Decreasing

- We fit the hurdle model and Poisson model to each data set
  - Model fits are compared based on the Deviance Information Criterion (DIC) and the Widely Applicable Information Criterion (WAIC)

Liver cancer

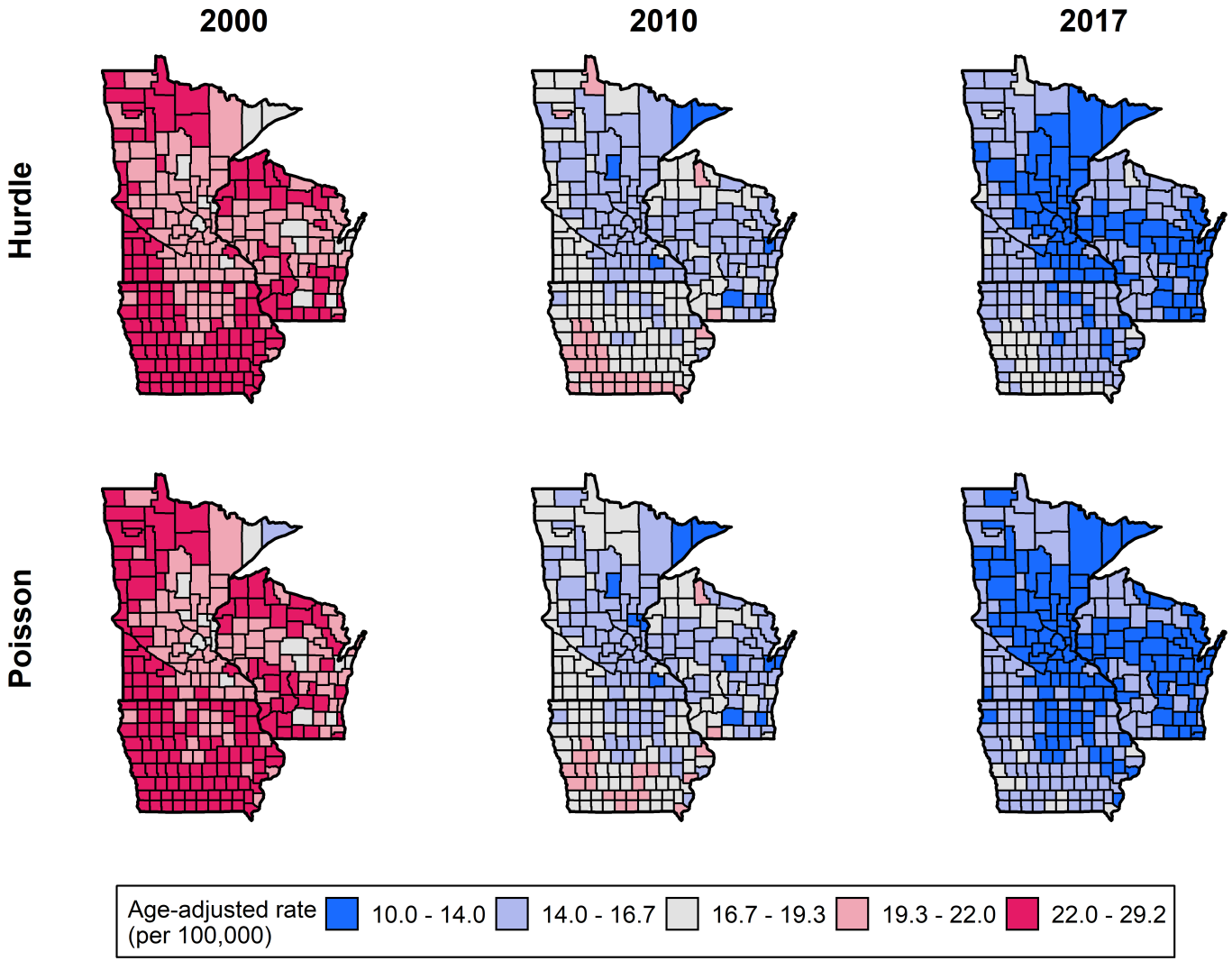


Liver cancer - comparison of approaches

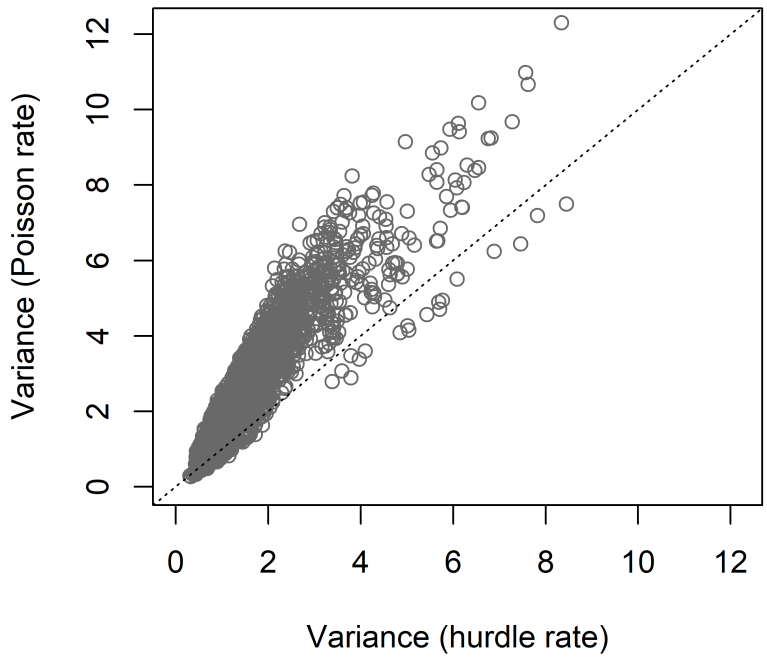


Cancer	Model	DIC	ΔDIC	WAIC	ΔWAIC
Liver	Hurdle	35,846	-16	35,853	-17
	Poisson	35,861		35,869	

Colorectal cancer



Colorectal cancer - comparison of approaches



Cancer	Model	DIC	ΔDIC	WAIC	ΔWAIC
Colorectal	Hurdle	89,007	41	89,013	39
	Poisson	88,966		88,975	

Simulation study

- **Goal:** Assess the performance of the hurdle model on simulated data sets with varying characteristics
- Specifically, we quantify the performance of the hurdle model compared to the Poisson model when fit to data that are truly hurdle-generated and under model misspecification (data are truly Poisson-generated)
- In the simulations, we utilize a simplified version of the proposed hurdle model (excludes population sizes in stage 1):

$$\begin{aligned} \text{logit}(\pi_{i,j,k}) &= \boldsymbol{x_k^T}\boldsymbol{\alpha} + \gamma_{1,i} + \delta_{1,j} \\ \log(\theta_{i,j,k}) &= \log(n_{i,j,k}) + \boldsymbol{x_k^T}\boldsymbol{\beta} + \gamma_{2,i} + \delta_{2,j} + \epsilon_{i,j} \end{aligned}$$

Simulation set-up

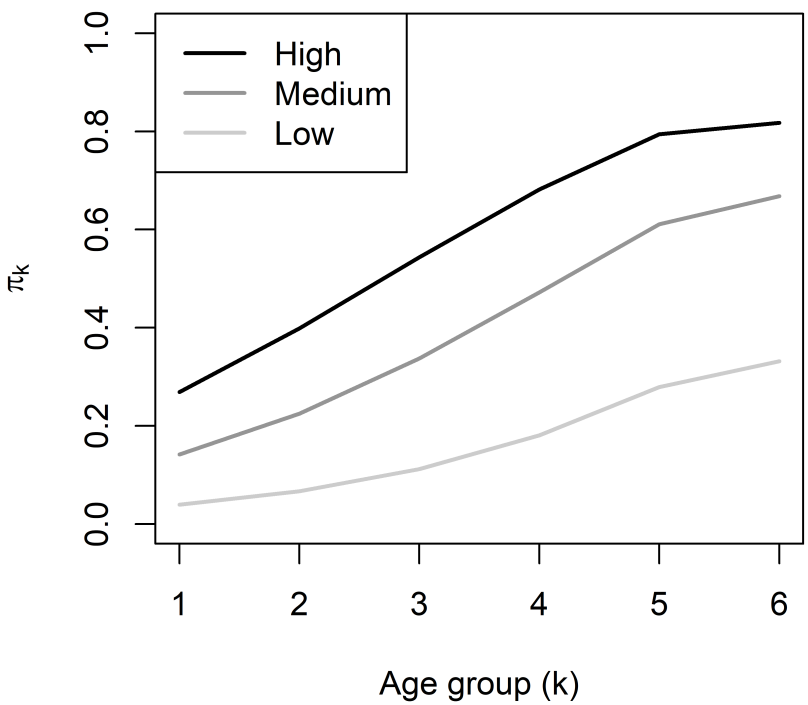
- Neighborhood structure and population sizes come from 44 counties in southern Minnesota from 2014-2017
- Individuals are classified into one of the following six age groups: $<$40, 40-49, 50-59, 60-69, 70-79, and 80$+$
- We use a proper CAR model to approximate the ICAR model
- Parameter values are set to be the following:

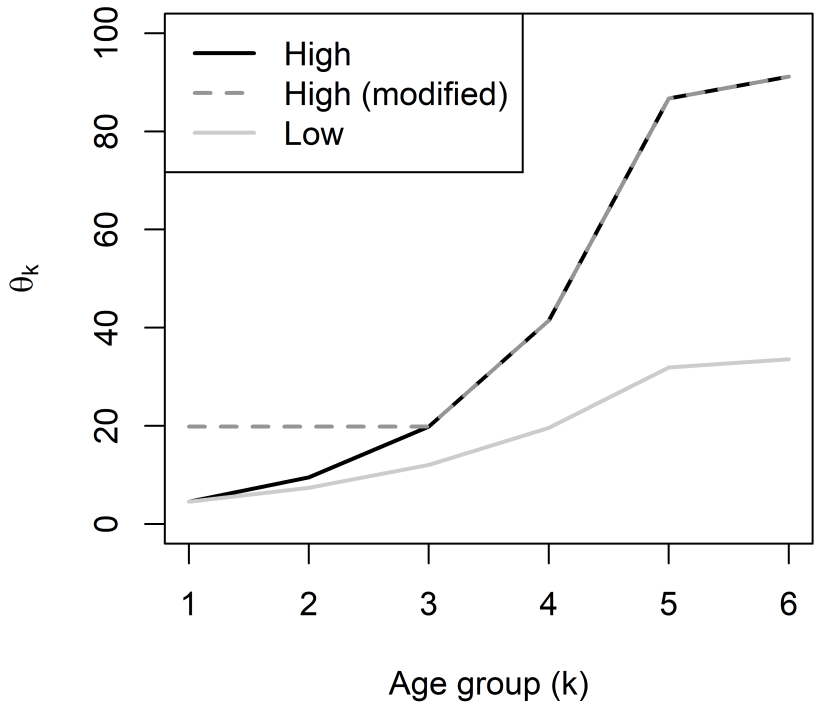
Parameter	**Value**
All precision terms	100
ρ_{γ_1}, ρ_{γ_2}, ρ_{γ}	0.9
ρ_{δ_1}, ρ_{δ_2}, ρ_{δ}	0.4

- We generate 100 data sets under each scenario

Simulation set-up (hurdle)

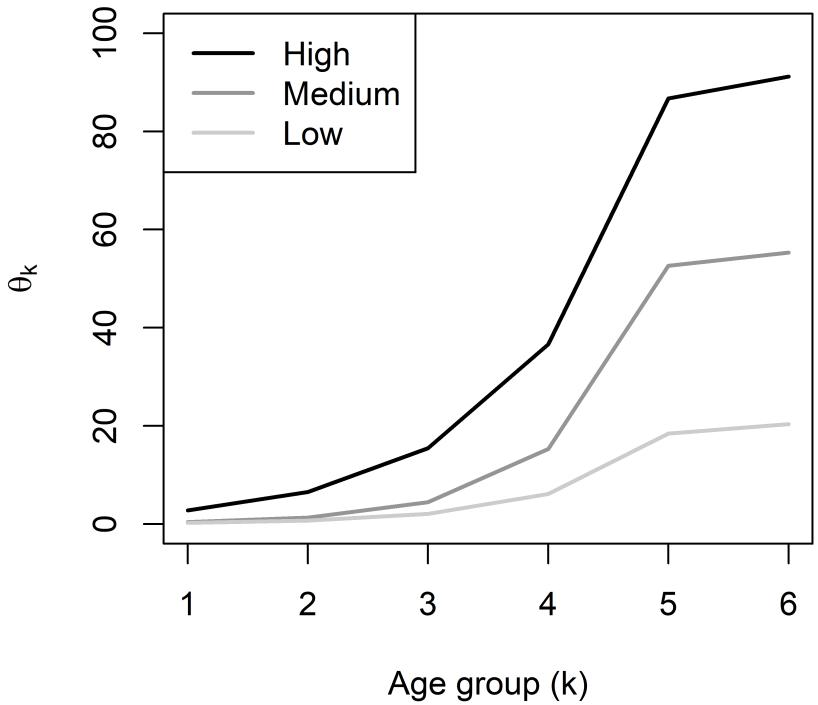
- In scenarios 1-9, data sets are generated from hurdle models
- $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are selected so that the average π_k and θ_k values are the following:





Simulation set-up (Poisson)

- In scenarios 10-12, data sets are generated from Poisson models with high, medium, and low count distributions
- $\boldsymbol{\beta}$ is selected so that the average θ_k values are the following:



Simulation results (hurdle-generated data)

Scenario	θ_k	π_k	Proportion zeros (Mean)	ΔDIC (Mean)	ΔDIC (SD)	Model fitting problems
1	Low	High	0.413	-685	82	0
2	Low	Medium	0.587	-657	91	0
3	Low	Low	0.831	-487	112	13
4	High	High	0.417	-1,093	157	0
5	High	Medium	0.591	-1,303	163	0
6	High	Low	0.832	-1,184	184	6
7	HM	High	0.415	-1,954	287	0
8	HM	Medium	0.590	-1,975	313	0
9	HM	Low	0.832	-1,540	302	1

- Changing the proportion of zeros had less of an effect on ΔDIC than changing the non-zero count distribution
- Model-fitting problems occurred when there was no variability in non-zero count values for at least one age group

Simulation results (Poisson-generated data)

Scenario	θ_k	Proportion zeros (Mean)	ΔDIC (Mean)	ΔDIC (SD)	Model fitting problems
10	High	0.440	45	10	0
11	Medium	0.638	42	9	0
12	Low	0.783	34	9	11

- The magnitude of ΔDIC was much smaller in scenarios 10-12 compared to scenarios 1-9
- This finding suggests that fitting a Poisson model to hurdle-generated data could have larger consequences than fitting a hurdle model to Poisson-generated data, in terms of DIC
- Model fitting problems occurred for the low values of θ_k due to a lack of variability in non-zero count values for at least one age group

Conclusions

- The Bayesian hierarchical hurdle model provides an improved fit to the Poisson model for the liver cancer data set but not for the colorectal cancer data set
  - Ultimately, the choice of a hurdle model or Poisson model is dependent on both the disease and the geographic region being studied
- Results from the simulation study suggest that the distribution of non-zero counts may be more influential in the hurdle model fit than the proportion of zeros in the data set
  - The hurdle model is likely to fit best on datasets where there are excess zeros but also high count values
  - In contrast, the Poisson model might adequately estimate a large proportion of zeros if the other count values are also low

Strengths and limitations of the proposed hurdle model

- Strengths
  - Accounts for excess zeros that occur for low-prevalence diseases and that occur as part of age group stratification in the modeling process
  - Assumes there is one zero-generating process, which often makes sense if everyone in the population is considered "at-risk"
  - In the examples, the hurdle model produced more precise estimates of age-adjusted rates than the Poisson model
- Limitations
  - More computationally expensive than the Poisson model
  - More complex model specification
  - Not possible to fit this model if there is no variability in the non-zero count values for a particular age group

Acknowledgements

- Environmental Health Sciences Research Center Pilot Grant No. NIEHS/NIH P30 ES005605
- National Science Foundation Graduate Research Fellowship Program Grant No. 000390183
  - Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation

References

1. Arab A. Spatial and spatio-temporal models for modeling epidemiological data with excess zeros. *Int J Environ Res Public Health*. 2015;12(9):10536–10548.
2. Banerjee S, Carlin BP, Gelfand AE. *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC; 2014.
3. Buescher PA. Age-adjusted death rates. Statistical Primer, No. 13. State Center for Health Statistics, North Carolina Divisionof Public Health; 2010.
4. Corpas-Burgos F, García-Donato G, Martinez-Beneito MA. Some findings on zero-inflated and hurdle poisson models for disease mapping. *Stat Med*. 2018;37(23):3325–3337.
5. Elliott P, Wartenberg D. Spatial epidemiology: current approaches and future challenges. *Environ Health Perspect*. 2004;112(9):998–1006.
6. Gelman A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal*. 2006;1(3):515-534.

References (continued)

7. Gómez-Rubio V. *Bayesian inference with INLA*. CRC Press; 2020.
8. Nandram B, Sedransk J, Pickle LW. Bayesian analysis and mapping of mortality rates for chronic obstructive pulmonary disease. *J Am Stat Assoc*. 2000;95(452):1110–1118.
9. National Center for Health Statistics. Detailed Mortality - All Counties (1968-2017), as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program.
10. Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J R Stat Soc Series B Stat Methodol*. 2009;71(2):319–392.
11. Watanabe S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J Mach Learn Res*. 2010;11(Dec):3571–3594.

Thank you for listening!

- Questions?
- Contact information:
  - Email: melissa-jay@uiowa.edu
  - Twitter: @MelissaJay