Epidemiology and Infection

Supplementary material

Quantifying differences in the epidemic curves from three influenza surveillance systems: a nonlinear regression analysis

E.G. Thomas[1†], J.M. McCaw[1,2], H.A. Kelly[3,4], K.A. Grant[3], J. McVernon[1,2]

† Corresponding author; emma.thomas@unimelb.edu.au

1. Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, University of Melbourne, Victoria, Australia
2. Vaccine and Immunisation Research Group, Murdoch Childrens' Research Institute, Victoria, Australia
3. Epidemiology Unit, Victorian Infectious Diseases Reference Laboratory, Victoria, Australia
4. National Centre for Epidemiology and Population Health, Australian National University, ACT, Australia

Methodological details: nonlinear regression

Here, we describe in greater detail the nonlinear regression model used to quantify the contribution of surveillance system, year, age group and region to the shape of the influenza epidemic curves. We had $N = 8,416$ response observations $y_1, \ldots, y_N$ representing counts of incident influenza cases in each surveillance week by year, surveillance system, age group and region. These attributes were encoded as categorical variables in the corresponding covariate observations $\mathbf{x_1}, \ldots, \mathbf{x_N}$.

We modelled the conditional mean number of incident cases in week t, $\lambda_t(\mathbf{x}, \beta)$, as a non-linear function $\lambda_t(\mathbf{x}, \beta) = f(t, \mathbf{x}, \beta)$ where Y_t is the disease count in week t, $\mathbf{x}$ is a vector of covariate values, β is a parameter vector and $\mathbf{x}$ is the covariate vector. We specified the function f to be the sum of a Gaussian function representing the epidemic component and a time-invariant constant representing the endemic component of weekly disease incidence, so that

$$f(t, \mathbf{x}, \beta) = \frac{S(\mathbf{x}, \beta)}{\sqrt{2\pi} D(\mathbf{x}, \beta)/3.92} \exp\left[\frac{(t - T(\mathbf{x}, \beta))^2}{2D(\mathbf{x}, \beta)^2}\right] + C(\mathbf{x}, \beta)$$

S may be interpreted as determining the final size of the epidemic component, D the length of the epidemic component, T the timing of the week of peak incidence and C the constant baseline or endemic activity in the system which is present year-round (both in and out of influenza season). As $D/3.92$ is the standard deviation parameter of the Gaussian function in equation (1) and T is its mean, we would expect approximately 95% of epidemic cases to occur in the time interval $T \pm 0.5D$, so that D approximates the duration of the seasonal epidemic; we refer to this quantity as the epidemic duration.

Setting $\beta^T = \left[\beta_{\mathbf{S}}{}^T \quad \beta_{\mathbf{D}}{}^T \quad \beta_{\mathbf{T}}{}^T \quad \beta_{\mathbf{C}}{}^T \quad \beta_{\alpha}{}^T\right]$, we then modelled $\log S$, D, T and $\log C$ as linear in $\mathbf{x}$ (where $\log a$ is the natural logarithm of a). We chose to model $\log S$ and $\log C$ rather than S and C as these parameters represent mean counts of disease and so are necessarily positive. Thus we defined four linear predictors:

$$\log S(\mathbf{x}, \beta) = \mathbf{x}^T \beta_{\mathbf{S}};\ D(\mathbf{x}, \beta) = \mathbf{x}^T \beta_{\mathbf{D}}; T(\mathbf{x}, \beta) = \mathbf{x}^T \beta_{\mathbf{T}};\ \log C(\mathbf{x}, \beta) = \mathbf{x}^T \beta_{\mathbf{C}} \tag{1}$$

For each linear predictor we also fit interaction terms between all variables and 'pandemic year' (2009) as an indicator variable. This decision was made a priori as we judged that the relative effects of surveillance system, age group and region were likely to be very different in a year in which a pandemic strain first appeared as compared to subsequent years, and evidence from the 2009 pandemic supports this.

We expected unmeasured sources of variation in the data to cause over-dispersion in the counts of incident disease. We therefore modelled the conditional weekly counts as independent negative binomial random variables $(Y_t|\mathbf{x}, \beta) \sim \mathrm{NB}\left(r(\mathbf{x}, \beta), p_t(\mathbf{x}, \beta)\right)$, where $p_t(\mathbf{x}, \beta) = \lambda_t(\mathbf{x}, \beta)/(r(\mathbf{x}, \beta) + \lambda_t(\mathbf{x}, \beta))$ and $\alpha(\mathbf{x}, \beta) = 1/r(\mathbf{x}, \beta)$ models the over-dispersion (larger values of α correspond to greater over-dispersion). We expected greater numbers of false positives in the ILI data, leading to additional sources of variability in case counts and thus greater over-dispersion of the ILI compared to the

laboratory confirmed data. We saw no a priori reason to expect over-dispersion to vary according to the other other variables in our model. We therefore allowed $\alpha(\mathbf{x}, \beta) = \mathbf{x}^T \beta_\alpha$ to vary by surveillance system.

For observations $y_1, \ldots, y_N$ and covariates $\mathbf{x_1}, \ldots, \mathbf{x_N}$, the log-likelihood for our model is given by

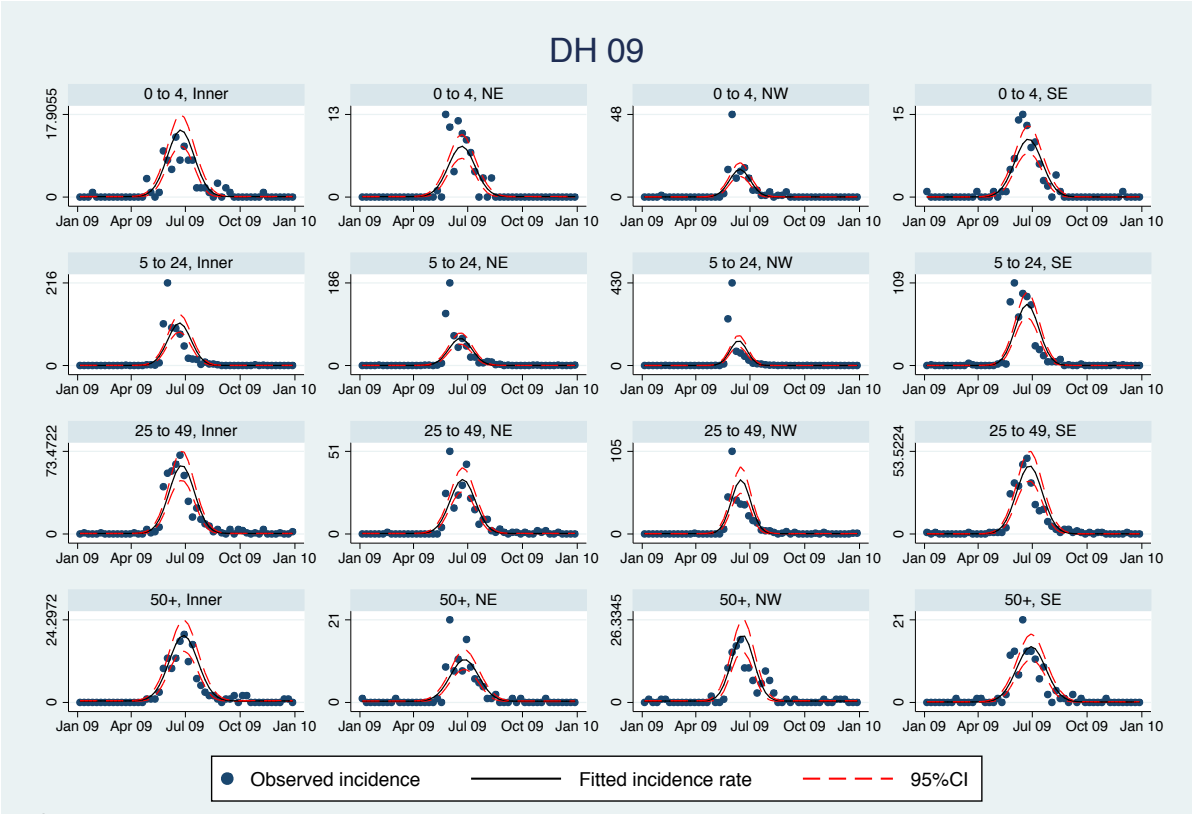
$$L = \sum_{i=1}^{N} r(\mathbf{x_i}, \beta) \log\left[1 - p_t(\mathbf{x_i}, \beta)\right] + y_i \log[p_t(\mathbf{x_i}, \beta)] + \log\Gamma\left[\frac{y_i + r(\mathbf{x_i}, \beta)}{(y_i + r(\mathbf{x_i}, \beta))r(\mathbf{x_i}, \beta)}\right]$$

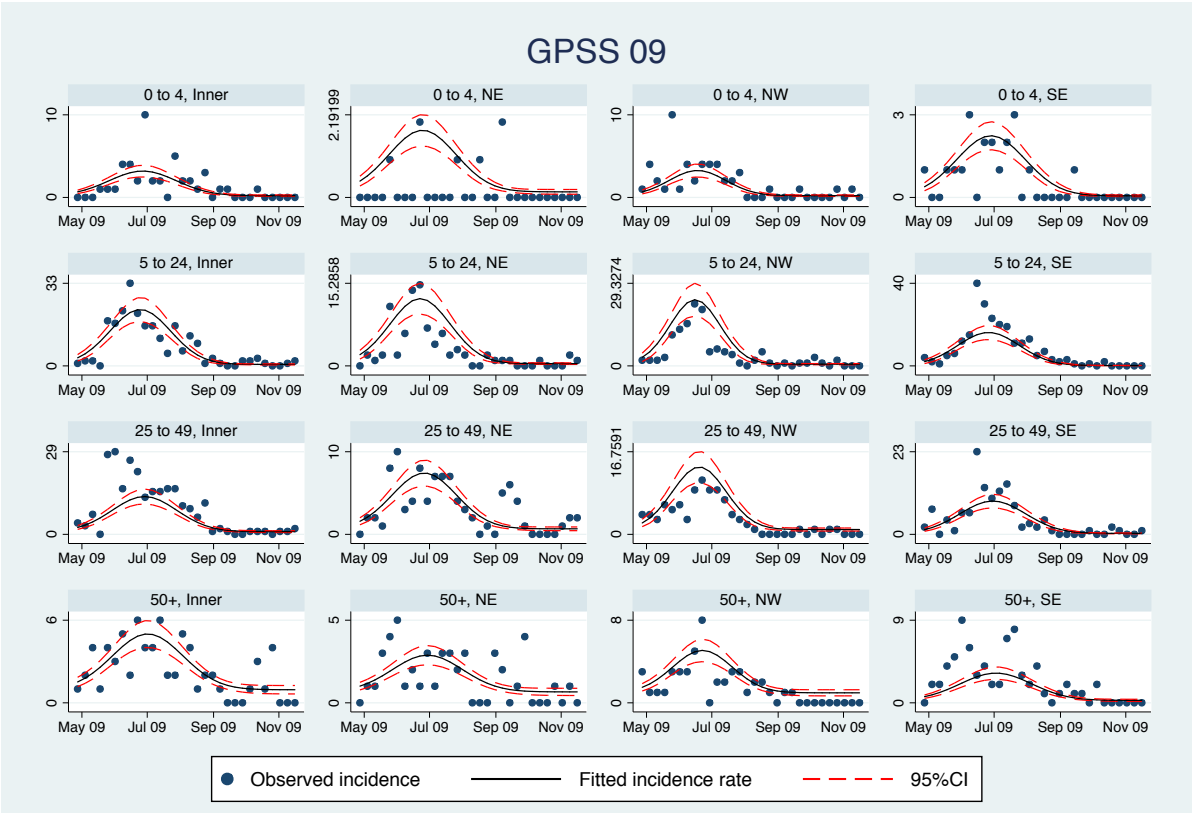
We numerically maximised L and computed likelihood-based confidence intervals for elements of β using the maximum likelihood estimation command `ml` in Stata.

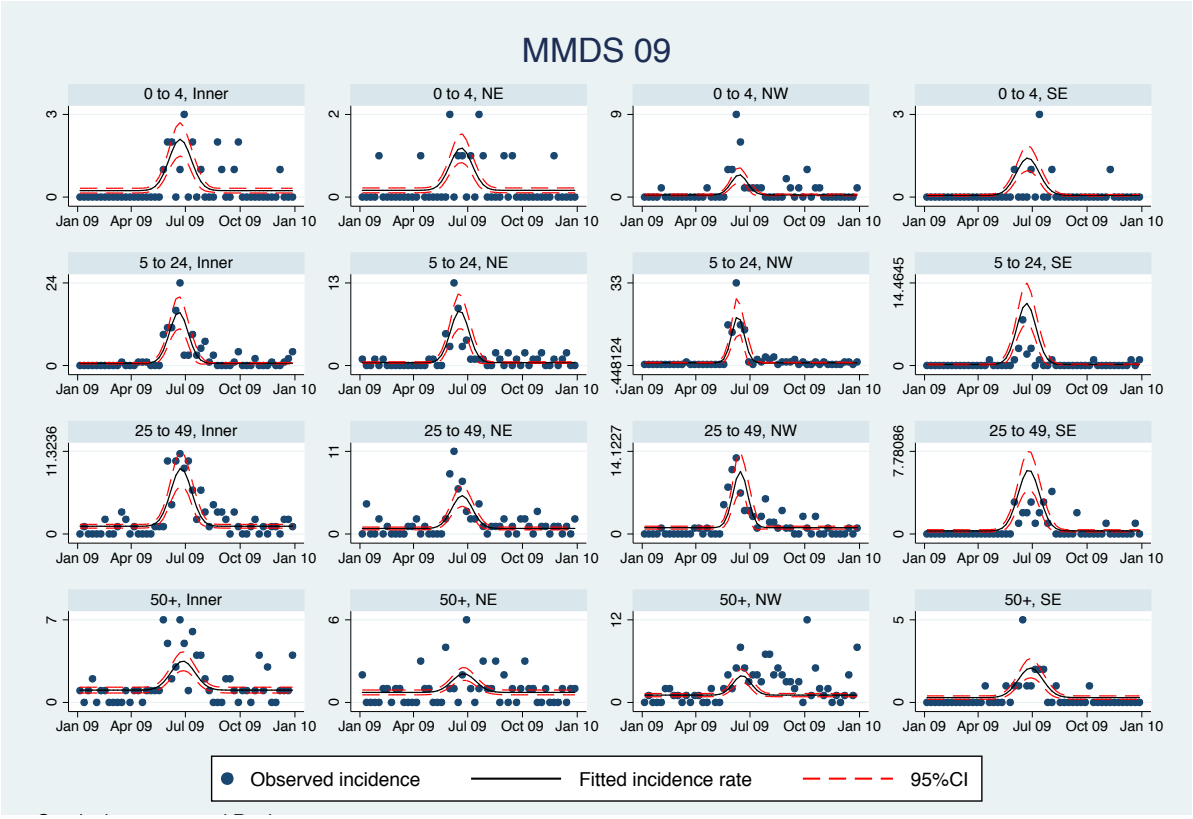
Data and nonlinear fits

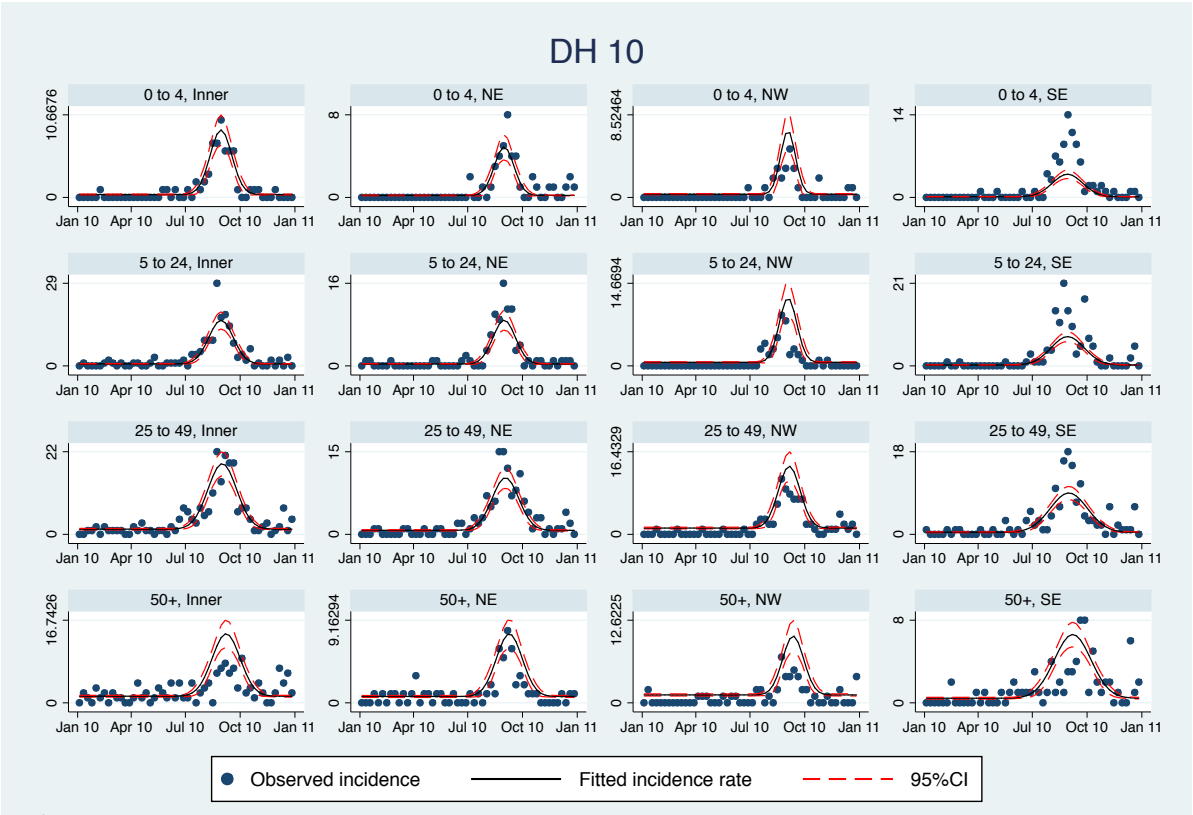
The following plots show the data and resulting fits using the nonlinear model described above. Each plot shows a time series of observed case counts for a particular year (2009–2012) surveillance system (Victorian Department of Health/DH, General Practice Sentinel Surveillance Scheme/GPSS, or Melbourne Medical Deputising Service/MMDS), age group (0–4, 5–24, 25–49 or 50+ years) and region of Melbourne (Inner, North West/NW, North East/NE or South East/SE). The nonlinear model for the mean disease count described by equation (1) and 95%CI is also shown.

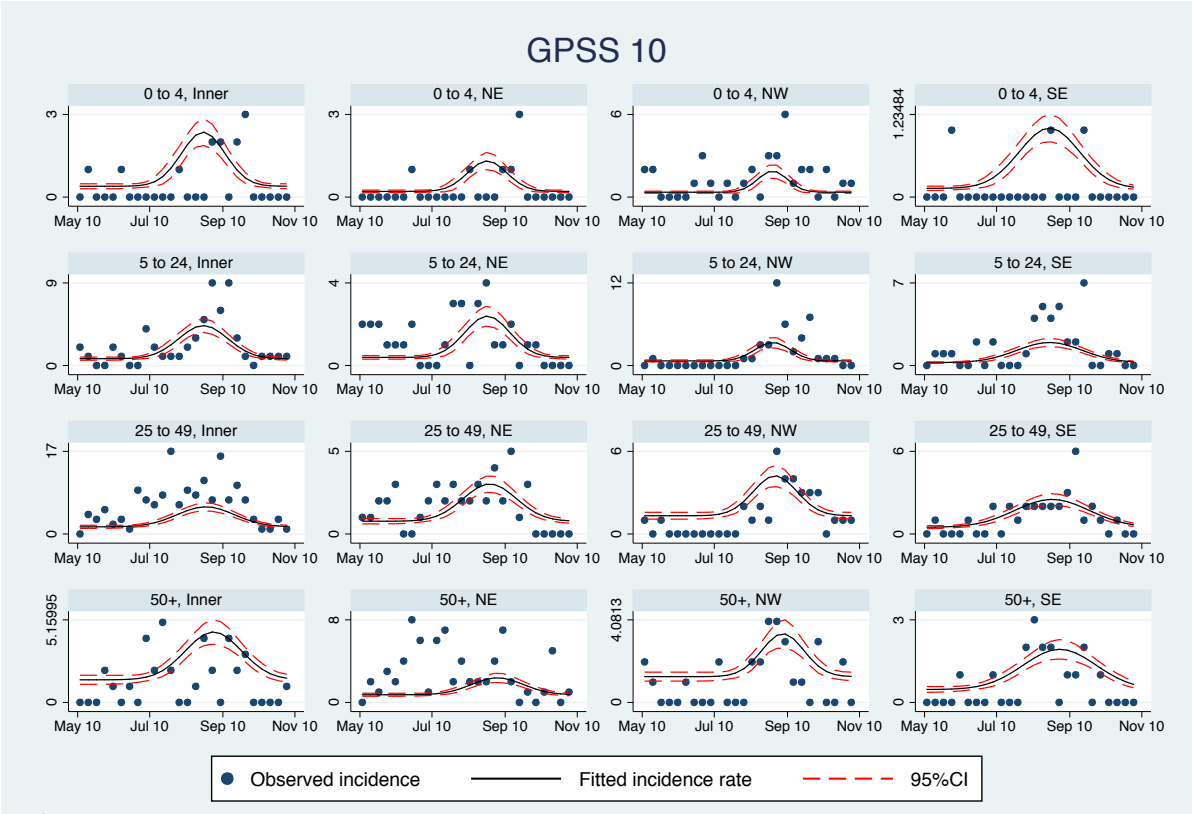
The graphs are organised by year and surveillance system. The relevant age group and region are shown above each subplot.

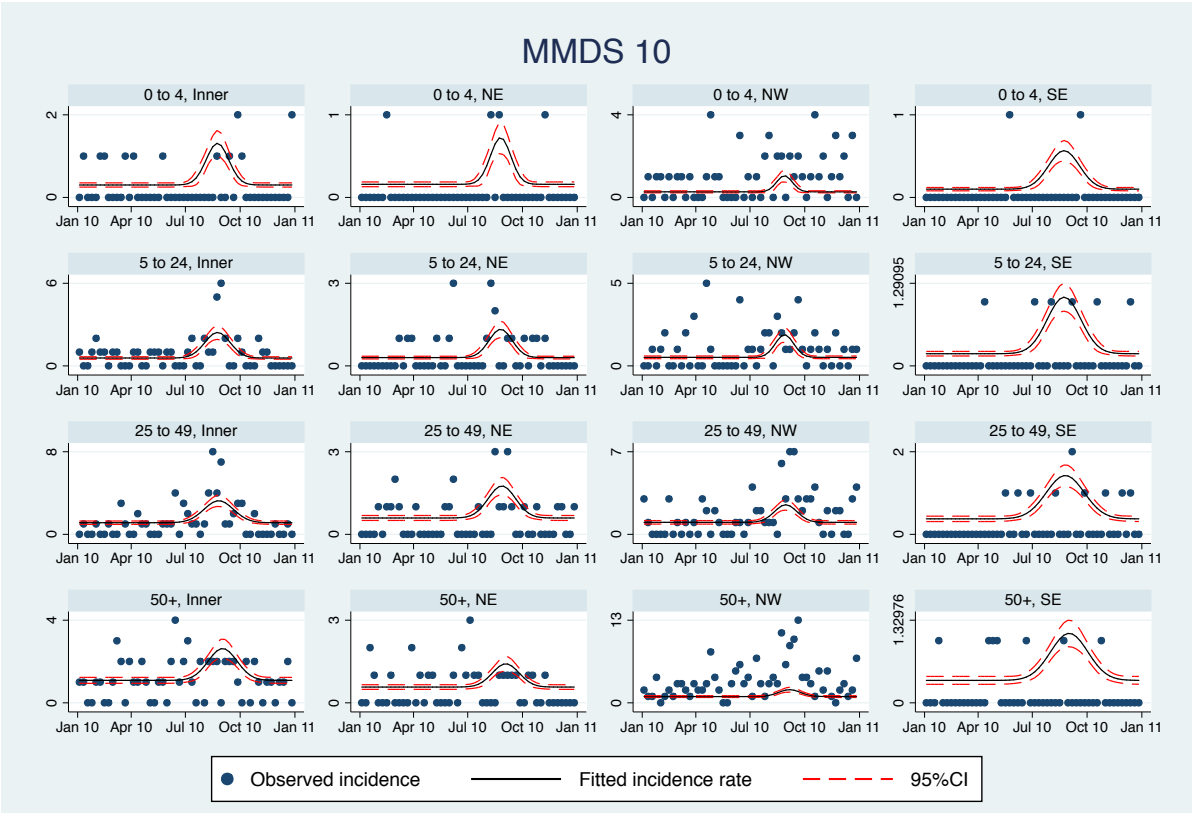

DH 09
0 to 4, Inner
0 to 4, NE
0 to 4, NW
0 to 4, SE
5 to 24, Inner
5 to 24, NE
5 to 24, NW
5 to 24, SE
25 to 49, Inner
25 to 49, NE
25 to 49, NW
25 to 49, SE
50+, Inner
50+, NE
50+, NW
50+, SE
Jan 09 Apr 09 Jul 09 Oct 09 Jan 10
Observed incidence
Fitted incidence rate
95%CI

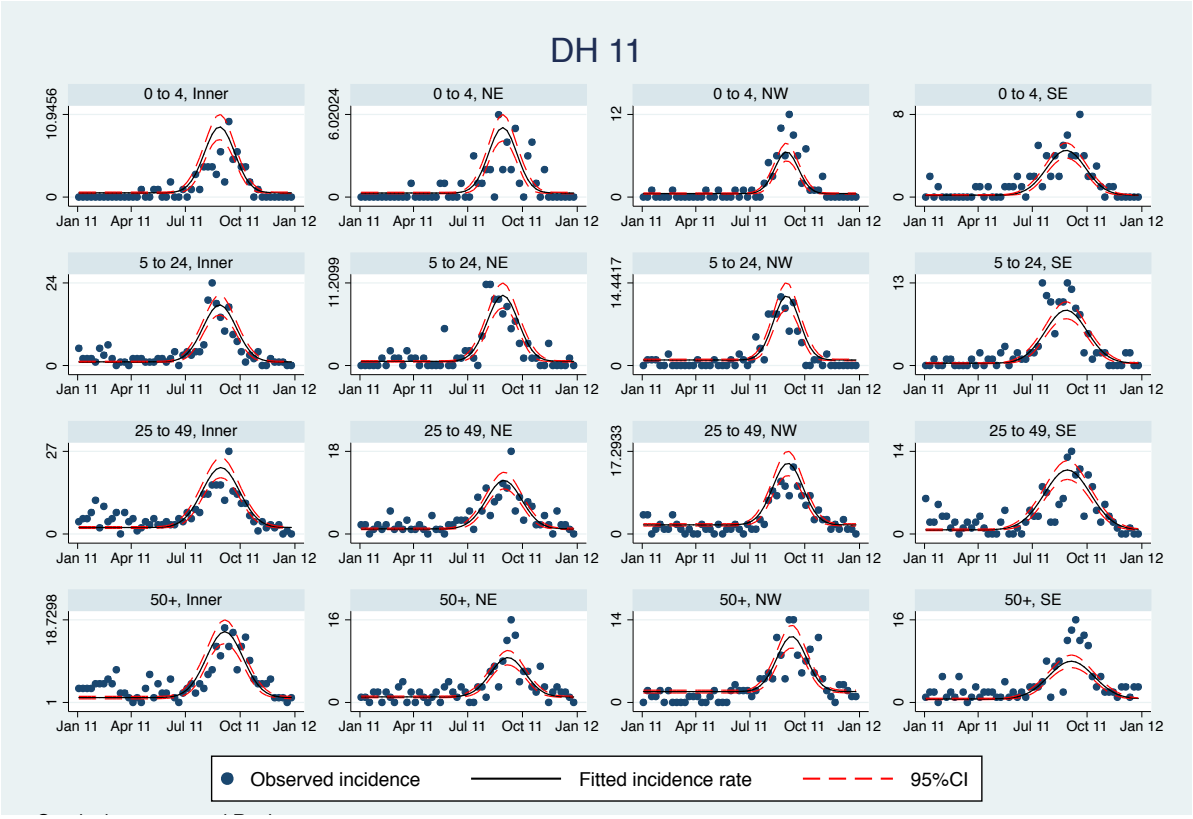

GPSS 09
0 to 4, Inner
0 to 4, NE
0 to 4, NW
0 to 4, SE
5 to 24, Inner
5 to 24, NE
5 to 24, NW
5 to 24, SE
25 to 49, Inner
25 to 49, NE
25 to 49, NW
25 to 49, SE
50+, Inner
50+, NE
50+, NW
50+, SE
May 09 Jul 09 Sep 09 Nov 09
Observed incidence
Fitted incidence rate
95%CI

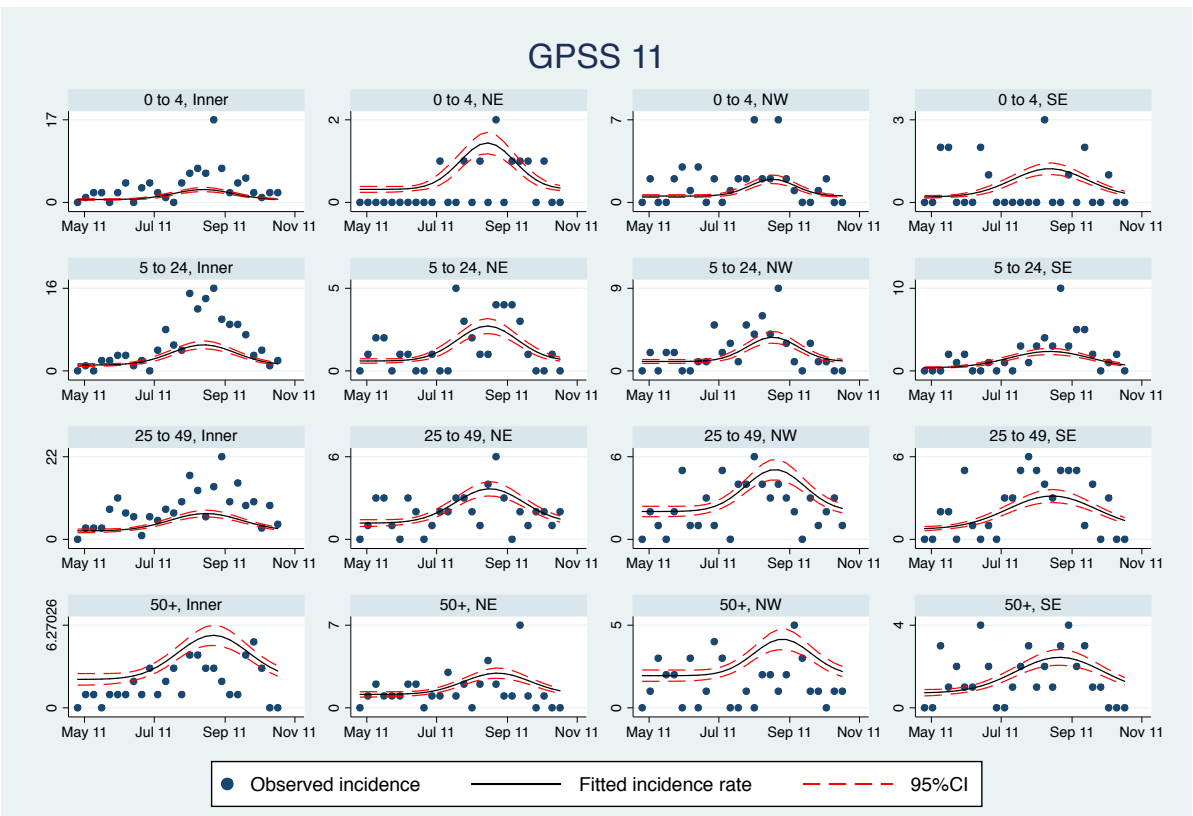

MMDS 09
0 to 4, Inner
0 to 4, NE
0 to 4, NW
0 to 4, SE
5 to 24, Inner
5 to 24, NE
5 to 24, NW
5 to 24, SE
25 to 49, Inner
25 to 49, NE
25 to 49, NW
25 to 49, SE
50+, Inner
50+, NE
50+, NW
50+, SE
Jan 09 Apr 09 Jul 09 Oct 09 Jan 10
Observed incidence
Fitted incidence rate
95%CI

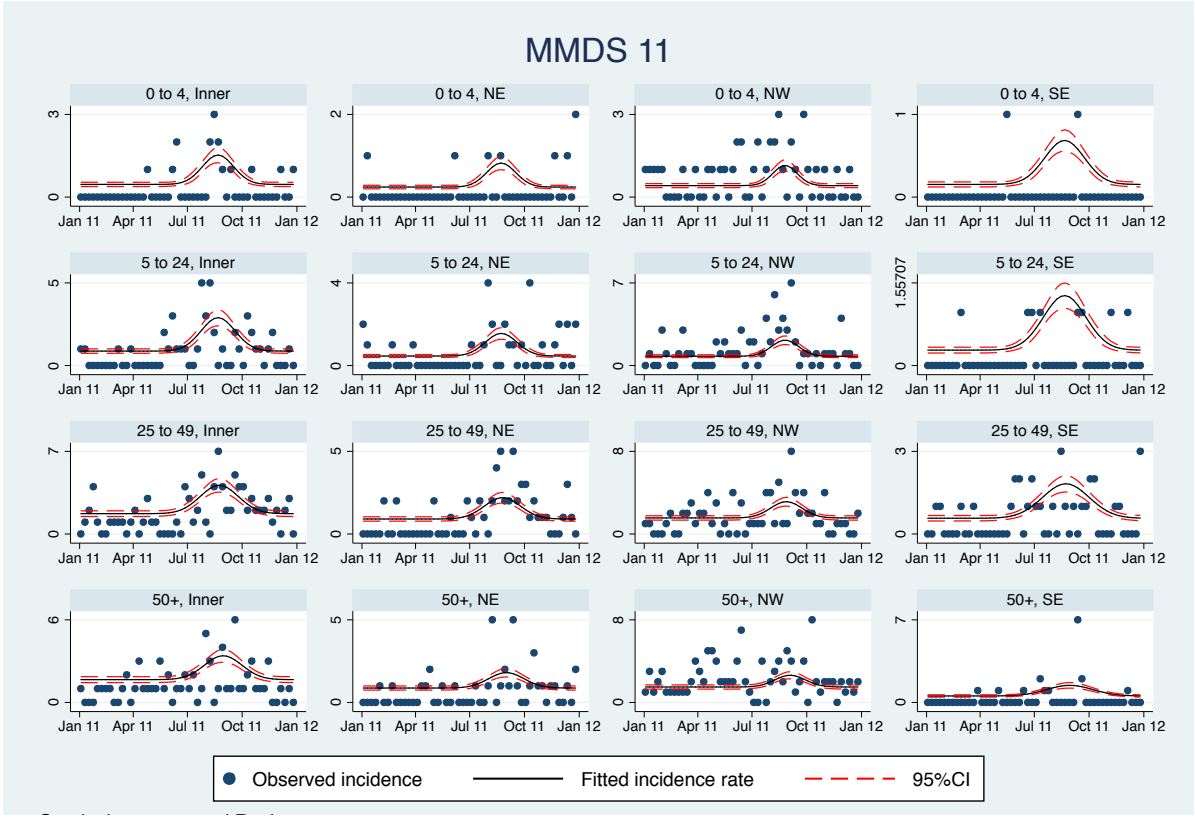

DH 10
0 to 4, Inner
0 to 4, NE
0 to 4, NW
0 to 4, SE
5 to 24, Inner
5 to 24, NE
5 to 24, NW
5 to 24, SE
25 to 49, Inner
25 to 49, NE
25 to 49, NW
25 to 49, SE
50+, Inner
50+, NE
50+, NW
50+, SE
Jan 10 Apr 10 Jul 10 Oct 10 Jan 11
Observed incidence
Fitted incidence rate
95%CI

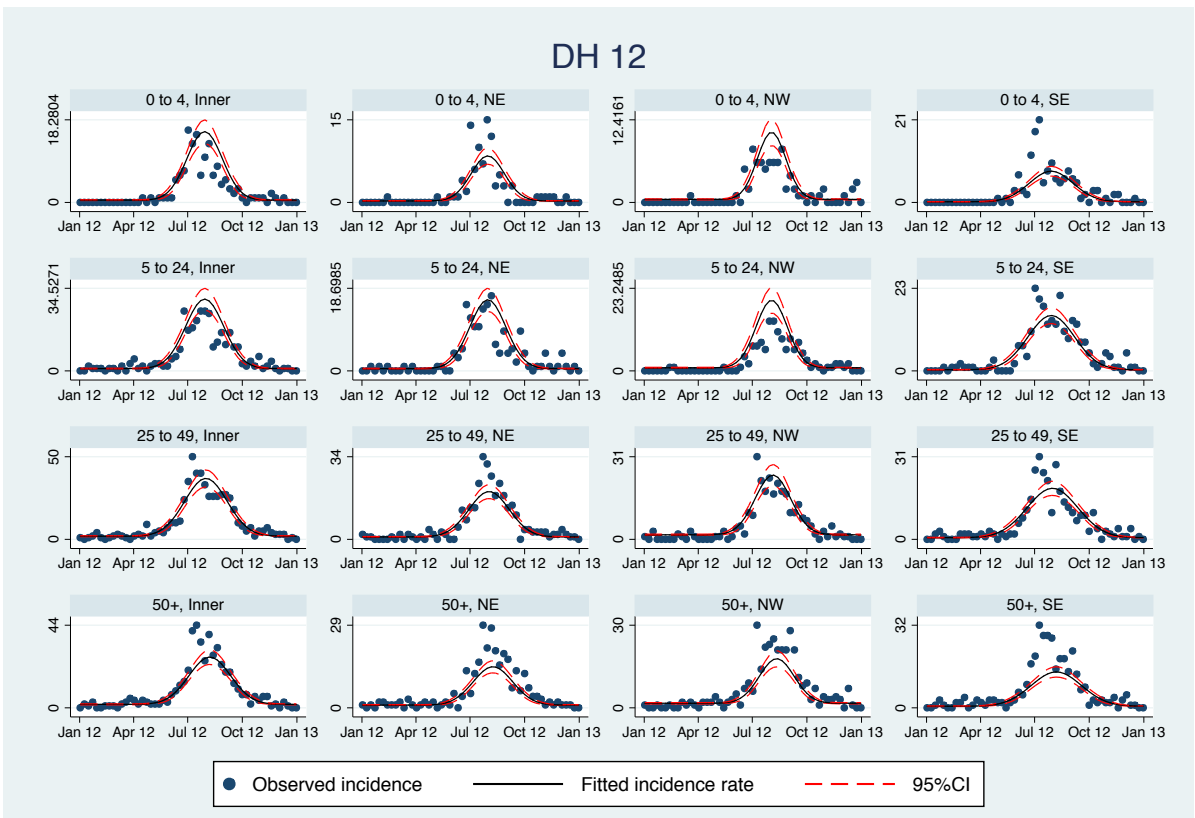

GPSS 10
0 to 4, Inner
0 to 4, NE
0 to 4, NW
0 to 4, SE
5 to 24, Inner
5 to 24, NE
5 to 24, NW
5 to 24, SE
25 to 49, Inner
25 to 49, NE
25 to 49, NW
25 to 49, SE
50+, Inner
50+, NE
50+, NW
50+, SE
May 10
Jul 10
Sep 10
Nov 10
Observed incidence
Fitted incidence rate
95%CI

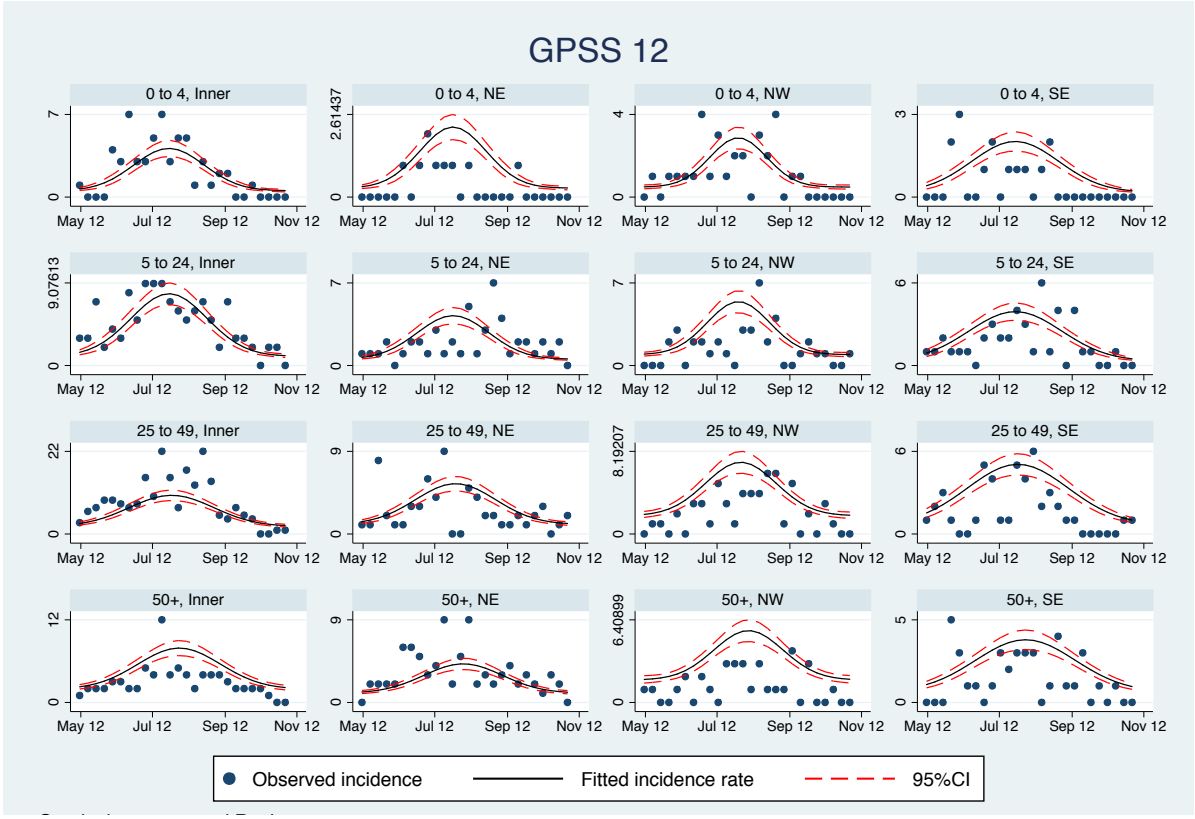

MMDS 10
0 to 4, Inner
0 to 4, NE
0 to 4, NW
0 to 4, SE
5 to 24, Inner
5 to 24, NE
5 to 24, NW
5 to 24, SE
25 to 49, Inner
25 to 49, NE
25 to 49, NW
25 to 49, SE
50+, Inner
50+, NE
50+, NW
50+, SE
Jan 10
Apr 10
Jul 10
Oct 10
Jan 11
Observed incidence
Fitted incidence rate
95%CI

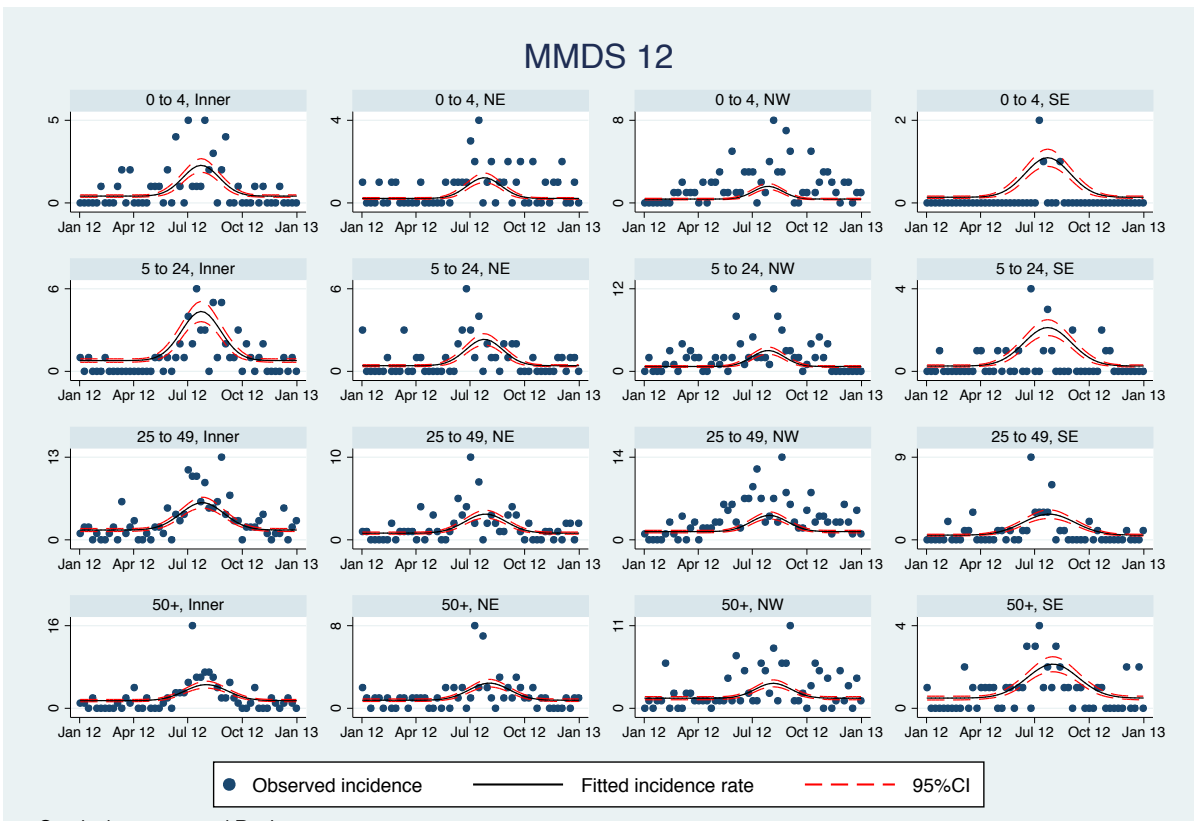

DH 11
0 to 4, Inner
0 to 4, NE
0 to 4, NW
0 to 4, SE
5 to 24, Inner
5 to 24, NE
5 to 24, NW
5 to 24, SE
25 to 49, Inner
25 to 49, NE
25 to 49, NW
25 to 49, SE
50+, Inner
50+, NE
50+, NW
50+, SE
Jan 11
Apr 11
Jul 11
Oct 11
Jan 12
Observed incidence
Fitted incidence rate
95%CI


GPSS 11
0 to 4, Inner
0 to 4, NE
0 to 4, NW
0 to 4, SE
5 to 24, Inner
5 to 24, NE
5 to 24, NW
5 to 24, SE
25 to 49, Inner
25 to 49, NE
25 to 49, NW
25 to 49, SE
50+, Inner
50+, NE
50+, NW
50+, SE
May 11
Jul 11
Sep 11
Nov 11
Observed incidence
Fitted incidence rate
95%CI


MMDS 11
0 to 4, Inner
0 to 4, NE
0 to 4, NW
0 to 4, SE
5 to 24, Inner
5 to 24, NE
5 to 24, NW
5 to 24, SE
25 to 49, Inner
25 to 49, NE
25 to 49, NW
25 to 49, SE
50+, Inner
50+, NE
50+, NW
50+, SE
Jan 11 Apr 11 Jul 11 Oct 11 Jan 12
Observed incidence
Fitted incidence rate
95%CI


DH 12
0 to 4, Inner
0 to 4, NE
0 to 4, NW
0 to 4, SE
5 to 24, Inner
5 to 24, NE
5 to 24, NW
5 to 24, SE
25 to 49, Inner
25 to 49, NE
25 to 49, NW
25 to 49, SE
50+, Inner
50+, NE
50+, NW
50+, SE
Jan 12 Apr 12 Jul 12 Oct 12 Jan 13
Observed incidence
Fitted incidence rate
95%CI


GPSS 12
0 to 4, Inner
0 to 4, NE
0 to 4, NW
0 to 4, SE
5 to 24, Inner
5 to 24, NE
5 to 24, NW
5 to 24, SE
25 to 49, Inner
25 to 49, NE
25 to 49, NW
25 to 49, SE
50+, Inner
50+, NE
50+, NW
50+, SE
May 12
Jul 12
Sep 12
Nov 12
Observed incidence
Fitted incidence rate
95%CI


MMDS 12
0 to 4, Inner
0 to 4, NE
0 to 4, NW
0 to 4, SE
5 to 24, Inner
5 to 24, NE
5 to 24, NW
5 to 24, SE
25 to 49, Inner
25 to 49, NE
25 to 49, NW
25 to 49, SE
50+, Inner
50+, NE
50+, NW
50+, SE
Jan 12
Apr 12
Jul 12
Oct 12
Jan 13
Observed incidence
Fitted incidence rate
95%CI