Spatial analysis of sporadic COVID-19 cases at the neighbourhood-level in Toronto, Ontario, 2020.

Lindsay Obress[1], Olaf Berke PhD[1], David N. Fisman, MD MPH[2], Ashleigh R. Tuite, PhD MPH[2], and Amy L. Greer, PhD[1, 2]

[1] Department of Population Medicine, University of Guelph

[2] Dalla Lana School of Public Health, University of Toronto

Corresponding Author: Amy L. Greer, PhD. Department of Population Medicine, University of Guelph. 50 Stone Rd E, Guelph, ON N1G 2W1. Email: agreer@uoguelph.ca

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Abstract

Background: As the largest city in Canada, Toronto has played an important role in the dynamics of COVID-19 transmission in Ontario. The burden of disease across Toronto neighbourhoods has shown significant heterogeneity. This study investigates spatial variation of sporadic COVID-19 cases in Toronto, Ontario and whether risk factors associated with socioeconomic status are related to the spatial variation.

Methods: A flexibly shaped spatial scan was used to detect clusters of increased risk of sporadic COVID-19 risk. Then, a generalized linear geostatistical model was used to investigate if average household size, population density, dependency ratio, and prevalence of low-income households were associated with sporadic COVID-19 rates.

Results: Three clusters of elevated COVID-19 risk were identified with standardized morbidity ratios ranging from 1.59–2.43. The generalized linear geostatistical model found that average household size (RR = 2.17, 95% CI: 1.80–2.61, $p < 0.01$) and percentage of low-income households (RR = 1.03, 95% CI: 1.02–1.04, $p < 0.01$) were significant predictors of sporadic COVID-19 cases at the neighbourhood-level.

Interpretation: Socioeconomic status is a well-established predictor of disease burden and may explain the spatial variation in sporadic COVID-19 cases across the Toronto neighbourhoods. Public policy that addresses the challenges faced by individuals in these communities are critical to curb the epidemic in Toronto and Canada as a whole.

Introduction

The first case of coronavirus disease 2019 (COVID-19) in Canada was reported on January 25th, 2020 after an individual returned to Toronto, Ontario from Wuhan, China (1). As the pandemic continued, Toronto has remained a focal area within Canada as the largest major city and home of Canada's busiest airport. As of November 26th, 2020, there have been 39,914 cases of COVID-19 reported in Toronto with a cumulative incidence of 1220.3 cases per 100,000 population (2,3). At that time, the cumulative incidence in the province of Ontario was 748.2 cases per 100,000 population (4). The risk and costs of a pandemic are not equal for all citizens. Individuals with low socioeconomic status disproportionally shoulder the burden of disease in any society, and this is amplified during a global health crisis (5). Lower socioeconomic status is associated with comorbidities linked to more severe COVID-19 disease and also with the conduct of essential work that cannot be done from home, such that these workers have continued to engage in in-person work throughout the pandemic (5,6). The spatial distribution of disease can provide insight into the observed differences in disease rates across a city by examining underlying social determinants of health and their relation to neighbourhood infection rates.

Toronto is subdivided into 140 neighbourhoods and the burden of COVID-19 has been observed to vary widely across the city (2). The goal of this study was to (i) determine if there are clusters of increased risk of sporadic COVID-19 at the neighbourhood-level, (ii) determine if there is spatial clustering in sporadic COVID-19 rates in Toronto, and (iii) create a generalized linear geostatistical model to investigate the effect of various risk factors on sporadic COVID-19 rates across Toronto.

Methods

Data sources

The COVID-19 case data was retrieved from the city of Toronto COVID-19 dashboard for cases reported between January 25th, 2020 – November 26th, 2020 (2). A case is defined as a confirmed or probable case of COVID-19 reported to Toronto Public Health through the Public Health Case and Contact Management Solution (CCM) (7). To explore the dynamics of spread at the community level, *sporadic* cases were selected, and outbreak related cases were excluded. The definition of sporadic cases is "*all cases not linked to an outbreak in general members of the population"* (7). The neighbourhood profiles and geographic boundary files were retrieved from Toronto Open Data (8,9). Case data and neighbourhood profiles, and geographical data files were linked by neighbourhood ID numbers. Population, average household size, population density, low-income measure – after tax (LIM-AT), percentage visible minority, and population size broken down by age group were selected from the 2016 Toronto Neighbourhood Profile as variables of interest (8). The variables of interest were selected as they are a subset of variables used to construct the Ontario Marginalization index, a widely used index that encompasses various factors of marginalization and socioeconomic status, however it is not available at the neighbourhood level (10). Population by age group was used to create a dependency ratio calculated as the ratio of children (<15 years old) and seniors (≥ 65 years old) to the population aged 15–64 for each neighbourhood (10).

Disease Mapping

The incidence rate of sporadic cases of COVID-19 reported from January 25th, 2020 to November 26th, 2020 in Toronto was mapped at the neighbourhood-level. Neighbourhood population size was used as the denominator to calculate the incidence rate for each neighbourhood. To account for varying population sizes across neighbourhoods, empirical Bayesian smoothed rates were estimated and their spatial distribution pattern was visualized by choropleth mapping (11). The UTM 17N projection was applied to minimize distortion of maps.

Disease cluster detection

A flexibly spatial scan test was used to determine the locations of probable geographic clusters of elevated sporadic COVID-19 rates and estimate the standardized morbidity ratio (SMR) within identified clusters (12). The flexibly spatial scan test was selected as it allows for irregularly shaped clusters to be detected that would not be picked up by more traditional methods (i.e., circular scanning window). The spatial scan test identifies clusters by gradually scanning each neighbourhood and increasing the scanning window to a maximum cluster size. The window that attains the maximum likelihood is identified as the primary, most likely, cluster. Additional clusters may then be identified. The maximum number of regions in a cluster was set to 14 as this represented 10% of neighbourhoods and the respective population would be still below the maximum size of 50% of total population for a single disease cluster. Identifying small clusters are preferred for public health studies to allow for intervention to be applied more easily, and clusters larger than 10-15% of the total regions are unlikely (12). P-values to determine significance of the spatial scan test were estimated using 999 Monte Carlo simulations, where the null hypothesis is that the rate of cases within a cluster does not differ from the rate outside of the cluster. The SMR was calculated by dividing the observed cases by the expected cases calculated in the flexibly shaped spatial scan test (12). We excluded clusters where the lower bound of the SMR 95% confidence interval was below 1.5, as spatial scan tests are most suitable to detect clusters with relative risk of 1.5 and above (13). Additionally, it was determined that a SMR above 1.5 would be of public health interest. Therefore, we excluded clusters with a SMR 95% confidence interval that was lower than 1.5.

Disease clustering

To determine if disease clustering (spatial dependence) was present in our data, two-sided Moran's I correlation coefficient was calculated using the empirical Bayesian smoothed rates, where the null hypothesis is absence of spatial correlation (14). Queen-neighbourhood structure was used for the test, where regions that share any border point are considered neighbours.

Generalized linear geostatistical model

To further investigate COVID-19 disease clustering in Toronto neighbourhoods, a model was built to examine risk factors. To account for spatial autocorrelation, a generalized linear geostatistical model (GLGM) was fit to model the effect of average household size, population density, LIM-AT, percentage visible minority, and dependency ratio on the number of sporadic COVID-19 cases at the neighbourhood level with population as the offset (15). The data is centered at the centroid of each neighbourhood and Euclidean distance was used to measure distances between neighbourhoods. The GLGM with a spherical spatial correlation structure with a Poisson family distribution was fit by Penalized-Quasi likelihood (PQL) estimation. The model was assessed by examining the normality assumption of the standardized residuals.

Analysis

R 4.0.2 was used to conduct all analyses including generating choropleth maps, flexible scan test (*smerc* package), spatial clustering tests (*spdep* package), and fitting GLGM (*MASS* and *GeoR* packages) (16,17). A significance level of 5% was used for all tests and confidence intervals.

Results

The dataset contained 30,598 sporadic cases of COVID-19 in Toronto across the 140 neighbourhoods. 2.3% of sporadic cases (704 cases) had missing postal codes and were excluded from the analyses. Reported laboratory confirmed case counts within a neighbourhood ranged from 27 to 1,115, with empirical Bayesian smoothed rates ranging from 263.8 to 3,367.8 cases per 100,000 population, and with a median of 823.5 cases per 100,000 population. Smoothed rates appear to be the highest in the north-west regions and north-east regions of the city and lowest in the southern and central regions (Figure 1).

The flexible scan test identified three regions of increased sporadic COVID-19 risk (Table 1; Figure 2). The primary cluster had the highest SMR of 2.43 (95% CI: 2.38–2.49), meaning there is a 2.43 times higher risk within this cluster compared to the risk of sporadic COVID-19 within the whole city of Toronto. The SMR of the secondary clusters were 1.59 (95% CI: 1.53–1.66) and 1.70 (95% CI: 1.59–1.82) (Table 1).

Moran's I test for clustering indicated that spatial clustering was present, indicating there is spatial dependence in the data that must be accounted for when modelling. The value of the Moran's I coefficient was 0.676 ($p < 0.01$).

A GLGM was fit and there was a significant effect of household size, and percentage of low-income households (defined by LIM-AT) on risk of sporadic COVID-19 cases. Population density, percentage visible minority, and the dependency ratio were not significant in the model and were removed. The final GLGM, including only average household size and percentage of low-income households, found both variables to be significant (Table 2). Where when average household size increases by 1, the increases risk of sporadic COVID-19 case by 2.17 ($\beta = 0.772$, RR = 2.17, $p < 0.01$), and a 1% increase in LIM-AT score increases risk of sporadic COVID-19 case by 1.03 ($\beta = 0.032$, RR = 1.03, $p < 0.01$) (Table 2). The range, the maximum distance between centroids of neighbourhoods up to which spatial dependence is observed by the model was 591 meters. The assumption of normality of residuals was found no violations.

Interpretation

Three clusters of elevated risk of sporadic COVID-19 cases were found within Toronto neighbourhoods with SMRs ranging from 1.59–2.43 (Table 1; Figure 2). While Cluster 1 is identified as the *most likely* cluster through the spatial scan test, all clusters are of importance for public health considerations. These clusters can be identified as key areas to target additional COVID-19 resources towards, such as pop-up testing clinics or targeted areas for vaccination.

The GLGM found that average household size and LIM-AT prevalence were associated with the rate of sporadic COVID-19 at the neighbourhood level (Table 2). For average household size, when the average household size in a neighbourhood increased by 1, the risk of sporadic COVID-19 increased by a factor of 2.17. Additionally, as the percentage of households that fall within the low-income measure criteria increased by 1%, the risk of sporadic COVID-19 cases increased by a factor of 1.03, at the neighbourhood level. Considering the difference between the neighbourhoods with lowest LIM-AT prevalence (4.5%) and the neighbourhoods with the highest prevalence (45.5%), there is a 3.67 times higher risk of sporadic COVID-19 for individuals living in the area with the highest LIM-AT prevalence. The model also had a low value for range, 591 metres. A range this low suggests that the spatial clustering can be explained by the risk factors included in the model and the identified clusters.

These findings align with literature linking poorer health outcomes to decreased socioeconomic status at a local level (5). A large-scale event such as a global pandemic only widens the discrepancies between those who are more and less privileged (5). Individuals who are of higher socio-economic status often work jobs where they can work from home more easily than those who are of lower socio-economic status (6). Those of lower socio-economic status often work in fields that have been deemed essential during a pandemic such as – healthcare, manufacturing, and retail, among others and may rely on public transit to get to their place of work (6,18). Policies are needed to address these risk factors and use information such as this to develop targeted strategies for vaccination. Paid sick days can prevent the spread of disease by giving working individuals the opportunity to seek medical care and isolate without lost wages or fear of termination, which is especially important for those in essential and low-income jobs (19–21). Additionally, providing locations, such as hotels, where individuals can safely isolate away from their families, may provide a solution for households that do not have space for at-home isolation, decreasing within-household spread (22,23).

This study is a first step into investigating the variability observed in the spatial distribution of SARS-CoV-2 cases during a pandemic. Further studies could examine additional factors that may better characterize socioeconomic status and marginalization. For example, using the Ontario Marginalization Index could be more representative of marginalization and socioeconomic status and can be constructed using census information, however, that was beyond the scope of this project (10). Individual level factors would also be of interest to examine, including occupations, ability to work from home, risk-taking behaviours, or children attending school in-person versus online. A separate research question could examine outbreak related cases such as in long-term care settings or school settings.

Limitations

Various types of bias may have been encountered when analysing these data. First, we are only looking at a limited set of group-level factors and summary values. This does not often give the full picture and may miss individual variation, such as specific sex, age, race differences, and additional variables may be of interest in future studies. Another limitation of this study is that only sporadic cases were investigated which could be influenced by misclassification bias. For example, individuals who work in a health care setting that test positive may be deemed part of an outbreak when their infection was acquired sporadically in the community or vice versa. There has also been found to be variation in testing rates across regions which may also influence the number of cases being detected in neighbourhoods. Additionally, when interpreting spatial studies, it is always important to consider the modifiable areal unit problem (MAUP) that occurs when studies aggregate spatial data to regions. The level of aggregation selected, in this study the neighbourhood level, effects the interpretation of the findings, as results may vary if another level of aggregation was selected (such as census tract or dissemination area). The flexibly shaped spatial scan test has limitations including being most practical for detection of small clusters and if larger clusters wanted to be considered, alternative methods would need to be used (12). These factors must be considered in the conclusions.

Conclusion

This study found wide variation in the spatial distribution of sporadic COVID-19 incidence rates across Toronto's 140 neighbourhoods. This variation can be at least partially explained by the risk factors that were considered in this study where residents of areas that have higher average household size and higher prevalence of low-income households had a higher risk of sporadic COVID-19. Policies such as paid sick days, hotel quarantine sites, and targeted vaccination strategies, could help close the gap in some of the inequalities identified in this study and could help prevent the spread of COVID-19.

References

1. Public Health Agency of Canada COVID-19 Surveillance and Epidemiology Team. A retrospective analysis of the start of the COVID-19 epidemic in Canada, January 15–March 12, 2020. Canada Commun Dis Rep. 2020;46(7/8):236–41.
2. City of Toronto. COVID-19: Status of Cases in Toronto [Internet]. City of Toronto. 2020 [cited 2020 Nov 28]. Available from: https://www.toronto.ca/home/covid-19/covid-19-latest-city-of-toronto-news/covid-19-status-of-cases-in-toronto/
3. Ontario Agency for Health Protection and Promotion (Public Health Ontario). Epidemiological summary: COVID-19 in Ontario - January 15, 2020 to November 26, 2020 [Internet]. Toronto, ON; 2020. Available from: https://covid-19.ontario.ca/covid-19-epidemiologic-summaries-public-health-ontario
4. Public Health Ontario. Ontario COVID-19 Data Tool [Internet]. 2021 [cited 2021 Mar 19]. Available from: https://www.publichealthontario.ca/en/data-and-analysis/infectious-disease/covid-19-data-surveillance/covid-19-data-tool?tab=summary
5. Singu S, Acharya A, Challagundla K, Byrareddy SN. Impact of Social Determinants of Health on the Emerging COVID-19 Pandemic in the United States. Front Public Heal. 2020;8(July):1–10.
6. Baker MG. Nonrelocatable Occupations at Increased Risk During Pandemics: United States, 2018. Am J Public Health. 2020;110(8):1126–32.
7. Toronto Public Health. External Dashboard Summary and Technical Notes [Internet]. Toronto, ON; 2020 [cited 2020 Dec 10]. p. 1–4. Available from: https://drive.google.com/file/d/1kq0d6sSLAFt2l8BUbnofn1-SrhBPREV6/view
8. City of Toronto. Neighbourhood Profiles [Internet]. Open Data Toronto. 2020 [cited 2020 Nov 28]. Available from: https://open.toronto.ca/dataset/neighbourhood-profiles/
9. City of Toronto. Neighbourhoods [Internet]. Open Data Toronto. 2020 [cited 2020 Nov 28]. Available from: https://open.toronto.ca/dataset/neighbourhoods/
10. Matheson FI, Moloney G, van Ingen T. 2016 Ontario Marginalization Index [Internet]. 2018. Available from: https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/download-
11. Berke O. Choropleth mapping of regional count data of Echinococcus multilocularis among red foxes in Lower Saxony, Germany. Prev Vet Med. 2001;52(2):119–31.
12. Tango T, Takahashi K. A flexibly shaped spatial scan statistic for detecting clusters. Int J Health Geogr. 2005;4(11):1–15.
13. Aamodt G, Samuelsen SO, Skrondal A. A simulation study of three methods for detecting disease clusters. Int J Health Geogr. 2006;5:1–11.
14. Assunção RM, Reis EA. A new proposal to adjust Moran's I for population density. Stat Med. 1999;18(16):2147–62.
15. Diggle P, Riberio PJ. Model-based Geostatistics. Springer; 2007. 79–98 p.
16. R Core Team. R: A language and environment for statistical computing. [Internet].

Vienna, Austria; 2020. Available from: https://www.r-project.org/%0A

17. RStudio Team. RStudio: Integrated Development Environment for R. [Internet]. Boston, MA; 2020. Available from: http://www.rstudio.com/

18. Sy KTL, Martinez ME, Rader B, White LF. Socioeconomic disparities in subway use and COVID-19 outcomes in New York City. Am J Epidemiol. 2020;

19. Vazquez J, Islam T, Beller J, Fiori K, Correa R, Correa DJ. Expanding Paid Sick Leave as a Public Health Tool in the Covid-19 Pandemic. J Occup Environ Med. 2020;62(10):e598–9.

20. Piper K, Youk A, James AE, Kumar S. Paid sick days and stay-At-home behavior for influenza. PLoS One. 2017;12(2):1–13.

21. Heymann J, Daku M. Ensuring equitable access to sick leave. CMAJ. 2014;186(13):975–6.

22. Jordan-Martin NC, Madad S, Alves L, Wang J, O'Gere L, Smith YG, et al. Isolation hotels: A community-based intervention to mitigate the spread of the COVID-19 pandemic. Heal Secur. 2020;18(5):377–82.

23. Cevik M, Baral SD, Crozier A, Cassell JA. Support for self-isolation is critical in covid-19 response. BMJ. 2021;372(224):1–2.

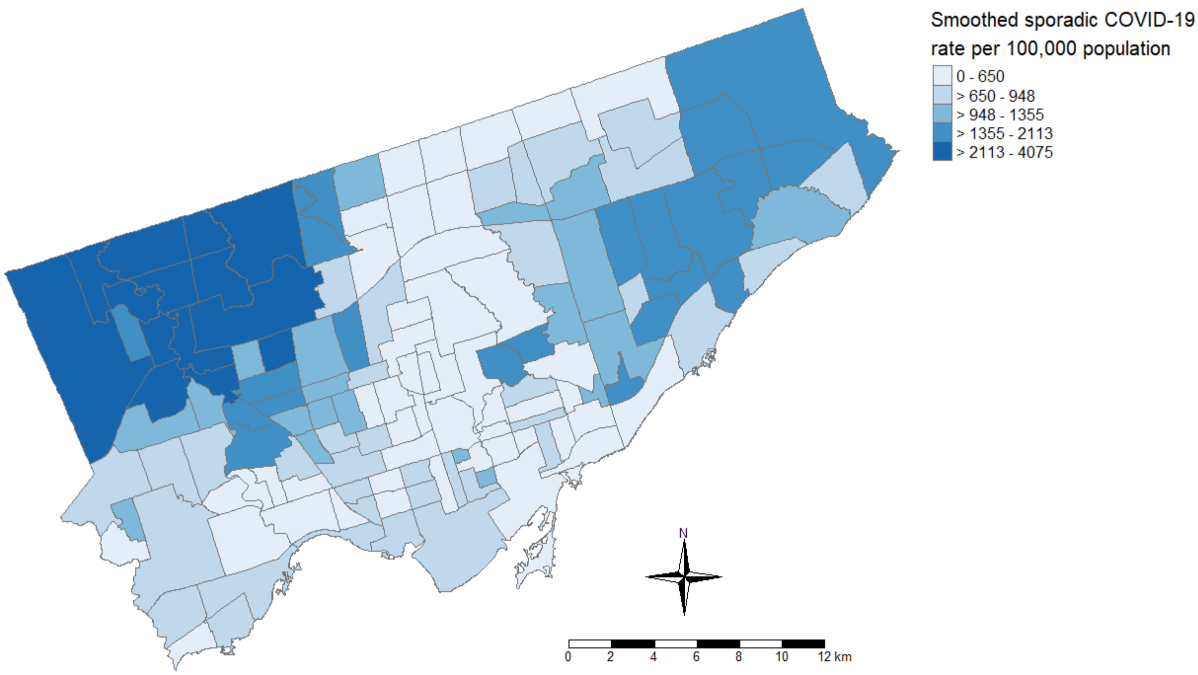
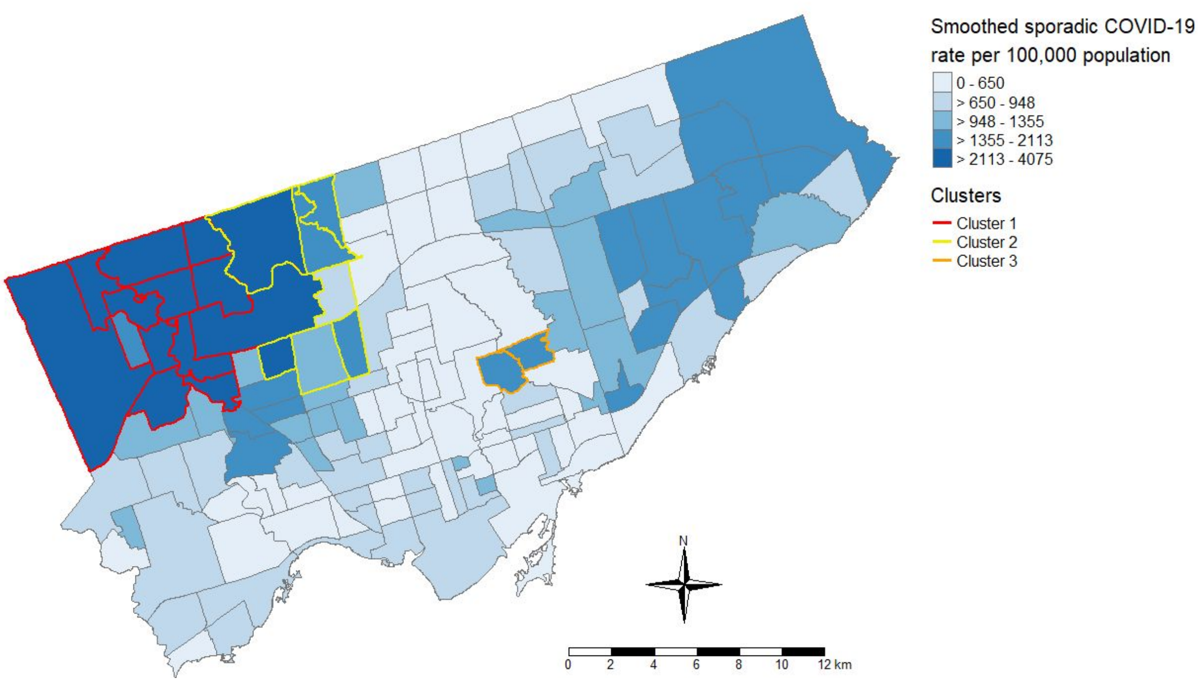


Figure 1: Map of smoothed sporadic COVID-19 cases per 100,000 population in Toronto, Ontario neighbourhoods from January 25, 2020 – November 26, 2020.



314 **Figure 2:** Three identified clusters of elevated sporadic COVID-19 risk in Toronto, ON.

315

316 **Table 1:** Clusters of increased risk of sporadic COVID-19 in Toronto neighbourhoods.

Cluster	SMR	95% CI	Cases	Expected	Population	P-value
1	2.43	2.38–2.49	6995	2873.49	262566	0.001
2	1.59	1.53–1.66	2323	1461.00	133499	0.001
3	1.70	1.59–1.82	802	471.04	43041	0.001

317

318

319

320

321

322 **Table 2:** Summary of generalized linear geostatistical model of sporadic COVID-19 cases in
323 Toronto, ON at the neighbourhood-level.

Variable	Parameter estimate	Standard Error	Relative risk (95% CI)	P-value
Intercept	–7.281	0.2493	0.0007 (0.0004 – 0.0011)	< 0.01
Average household size	0.772	0.0951	2.17 (1.80–2.61)	< 0.01
LIM-AT	0.032	0.0048	1.03 (1.02–1.04)	< 0.01
Range	0.591	–	–	–

324

Appendix

Table A1: Summary statistics of Toronto neighbourhood characteristics.

	Median	Range
Case Count	137	27–1115
Population	16750	6577–65913
Raw Rate per 100,000	819	255–3384
Smoothed Rate per 10,000	824	264–3368
Average Household Size	2.525	1.540–3.44
LIM-AT	18.55	4.50–45.50
Dependency Ratio	0.4683	0.1291–0.6980
Population density per km^2	5072	1040–44321