SPATIAL MODELLING OF COVID-19 INCIDENCE RATE IN CANADA

Sarah N. Fatholahi [1], Charlotte Pan [1], Lanying Wang [1], Jonathan Li *[1]

[1] Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON N2L 3G1, Canada

Commission IV, WG IV/3

KEY WORDS: Statistical Regression, COVID-19, Ordinary Least Squares, Spatial Error Model, Spatial Lag Model, Canada.

ABSTRACT:

The COVID-19 was first declared by World Health Organization (WHO) as global pandemic on March 11th 2020. While most of COVID-related studies have focused on epidemiological perspective, the spatial analysis of disease outbreak is also important to provide perceptions of transmission rates. Therefore, this paper attempts to identify the potential factors contributing to the COVID-19 incidence rate at provincial-level in Canada. Three statistical regression models, ordinary least squares (OLS), spatial error model, and spatial lag model (SLM) were applied to 14 independent variables including socio-demographic, economic, weather, health and facilities related factors. The results indicated that three factors including median income, diabetes and unemployment significantly affected the COVID-19 rates in Canada. Among three global models, the SLM performed the best to explain the key variables and spatial variability of disease incidence with a R^2 value of 61%. However, in this study, the application of local regression models such as geographically weighted regression (GWR) and multiscale GWR (MGWR) have not been considered and this could be a scope for the future research.

1. INTRODUCTION

The ongoing COVID-19 was first declared by World Health Organization (WHO) as global pandemic on March 11th 2020 (WHO, 2020). The United Nations declared this pandemic as a world crisis as most of developing nations experienced its severe burden on their national economy. However, the negative impacts of COVID-19 are not limited to the developing nations. Moreover, the demographic factors such as migrants, aging and population play a significant role in spreading and shaping the COVID-19's pattern across the world. For instance, more than 95% of people who have died in Europe due to this pandemic have been over 60 years old (WHO, 2020).

While most of COVID-related studies have focused on epidemiological perspective, the spatial analysis of disease outbreak is also important to provide perceptions of transmission rates. The development of geospatial techniques and GIS have enabled the analysing and visualizing of disease transmission (Zhou et al., 2020).

The provided spatial information can support for decision-making in disaster management and spread of epidemics. Implementing and fine-tunning GIS-based modelling techniques to analyse the spatiotemporal patterns between COVID-19 growth rate and variations among the socio-economic, environmental, and demographic sector data can be a great asset to the current pandemic research field, especially to better explain and forecast oscillations in the COVID-19 outbreak (Mollalo et al., 2020; Hassan et al., 2020).

To name a few of the most recent researches, a study applying spatial regression models to the European region found that demographic factor such as total population and the percentage of the elderly population (age 65+) are strong contributors to both COVID-19 cases and deaths (Sannigrahi et al., 2020). A global study applying geographically weighted regression (GWR) model further confirmed that the percentage of teenagers and adults (age between 15 to 64), among the population has a strong positive correlation with COVID-19 cases (Rex et al., 2020). Likewise, many researchers using spatial modelling techniques have reached meaningful conclusions about the roles of socio-economic factors in the current pandemic. For instance, Rahman et al. (2021) showed that there is a strong association between COVID-19 incidence rates in Bangladesh and economic factors including monthly consumption, the number of health workers, and distance from the capital city. Rex et al. (2020) suggested that the out-of-pocket expenditure can also significantly affect the COVID-19 at the national level for 175 countries. Mollalo et al. (2020) selected four variables from a total of thirty five independent variables including demographic, socioeconomic, environmental, and topographic to implement spatial regression modelling. Zhang and Schwartz (2020) applied a global spatial regression model to investigate the spread of virus in both metropolitan and non-metropolitan regions.

Spatial modelling is considered as an effective tool for statistically and geographically analysing the relationship between disease transmission rate and some explanatory variables (Dickin et al. 2013; Neil, 2006; Pickle, 2002
). In this research, three global spatial models are examined to investigate how well the variations of COVID-19 can be explained based on demographic, socioeconomic, meteorological and health-related factors as explanatory variables affecting disease infection rate in Canada. To the best of our knowledge, this is the first attempt to utilize geographic analysis of COVID-19 outbreak in Canada to provide valuable information for policy making.

*Corresponding author: Jonathan Li (junli@uwaterloo.ca)

2. MATERIAL AND METHOD

2.1 Dataset

The first cases of COVID-19 in Canada was reported on Jan 25, 2020 and represented a significant challenge for public health and health care systems. The COVID-19 related data applied in this study was collected from Johns Hopkins University, with the number of confirmed and death cases until May 30, 2021 (Figure 1). A total of fourteen demographic, socioeconomic, meteorological, and health-related factors have been considered as explanatory variables in modelling process. The description of the selected parameters is given in Table 1 (Statistics Canada). These variables have been prepared at the provincial-level, and collected in ArcGIS environment. Three different models were applied to investigate the relationship between COVID-19 incidence rate (as dependent variable) and the potential explanatory (independent) variables. The models are ordinary least squares (OLS), spatial lag model (SLM), and spatial error model (SEM). The OLS model was performed using ArcMap 10.8.1. Both SLM and SEM models were implemented in GeoDa software (The University of Chicago, Chicago, IL, USA).

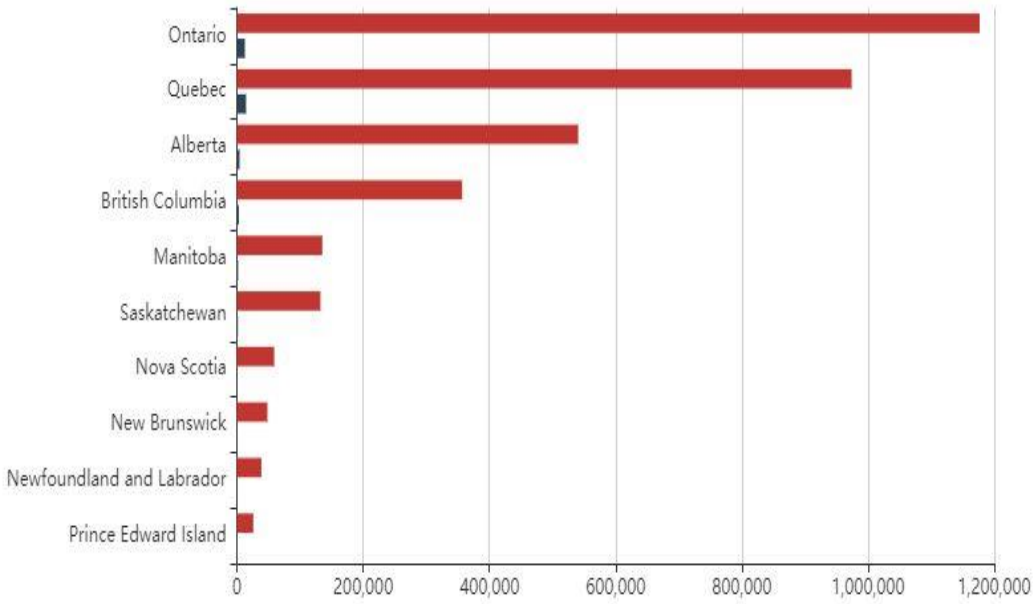

Figure 1. Confirmed cases by Canadian provinces

2.2 Ordinary least squares (OLS)

The spatial regression models have been widely used to examine demographic analysis and environmental monitoring (Chi & Zhu, 2008; Jain et al. 2019; Fang et al. 2015). The main application of spatial regression models is to understand the spatial states of a feature distribution such as autocorrelation, stationarity and heterogeneity. In this research, the total three global spatial regression models including Ordinary Least Square (OLS), Spatial Lag Model (SLM) and Spatial Error Model (SEM) have been carried out to analyse how various environmental, demographic, socioeconomic, and health factors can contribute to COVID-19 transmission rate in Canada. The general framework for this research is represented in Figure 2. The OLS is a linear technique used to regress a response variable (COVID-19 incidence rate) on a set of predictors or independent variables. OLS regression model is a powerful method for modelling and predicting continuous data, especially when it is applied in conjunction with data transformation (Hutcheson, 2011). OLS assumption is based on homogeneity, spatial non-variability and a constant relationship over space (Mollalo et al. 2020). This model is calculated as follows:

$$y_i = b_0 + x_i b + e_i \quad (1)$$

where y_i indicates the COVID-19 incidence rate at the *i*th location; b_0 is the intercept and indicator of the value of y when x is zero; b denotes the vector of regression coefficients describing the changes of y; and e_i is the random error. The basic function for OLS model is optimizing the coefficients by decreasing the sum of squared errors (Oshan et al., 2019). This model is based on the assumption that the residual errors are homogenous, therefore it is not efficient enough when these errors are spatially correlated and therefore will lead to a bias in estimating regression coefficients.

2.3 Spatial lag model (SLM)

The SLM and SEM are the two regression spatial models which are the variants of OLS model (Anselin, 2003; Ward and Gleditsch, 2018). The SLM is used to reflect the impact of spatial units on other units in the area, in which the spatial lag in dependent variable is considered (Wang et al. 2017). The SLM considers dependency between the explanatory variables.

It also assumes a close association between the dependent and a set of independent variables, and is characterized by:

$$y_i = b_0 + x_i\, b + \rho\, w_i + e_i \qquad (2)$$

where ρ is the spatial autoregressive coefficient; w is the spatial weight matrix representing distance relationships between observations (Mollalo et al., 2020). In fact, the spatial lag function which evaluates the impact of adjacent variables on each other, can be utilized as an independent variable as well in the modelling process (Wu et al., 2020).

2.4 Spatial error model (SEM)

The SEM can deal with the problem of spatial autocorrelation with independent error term (Izon et al. 2016). The SEM considers the spatial dependency in the error term of OLS model, and divides the error into two spatial components of error term and random error ($\lambda w_i\, \xi_i$ and e_i) as follows:

$$y_i = b_0 + x_i\, b + \lambda w_i\, \xi_i + e_i \qquad (3)$$

where ξ_i is the spatial component of the error; λ represents the correlation levels between the components; and e_i is the spatial uncorrelated error (Ward and Gleditsch, 2018).

Theme	Independent variables
Demographic	1. Total population 2. Population of over 65 years 3. Population density 4. International migration rate
Socioeconomic	5. Median household income 6. Gross Domestic Product (GDP) 7. Unemployment rate 8. Poverty rate
Meteorologic	9. Average temperature 10. Annual precipitation
Health-related	11. Diabetes rate 12. Adult smoking 13. Total number of hospitals 14. Total number of nurses

Table 1. Description of explanatory variables used in this study.

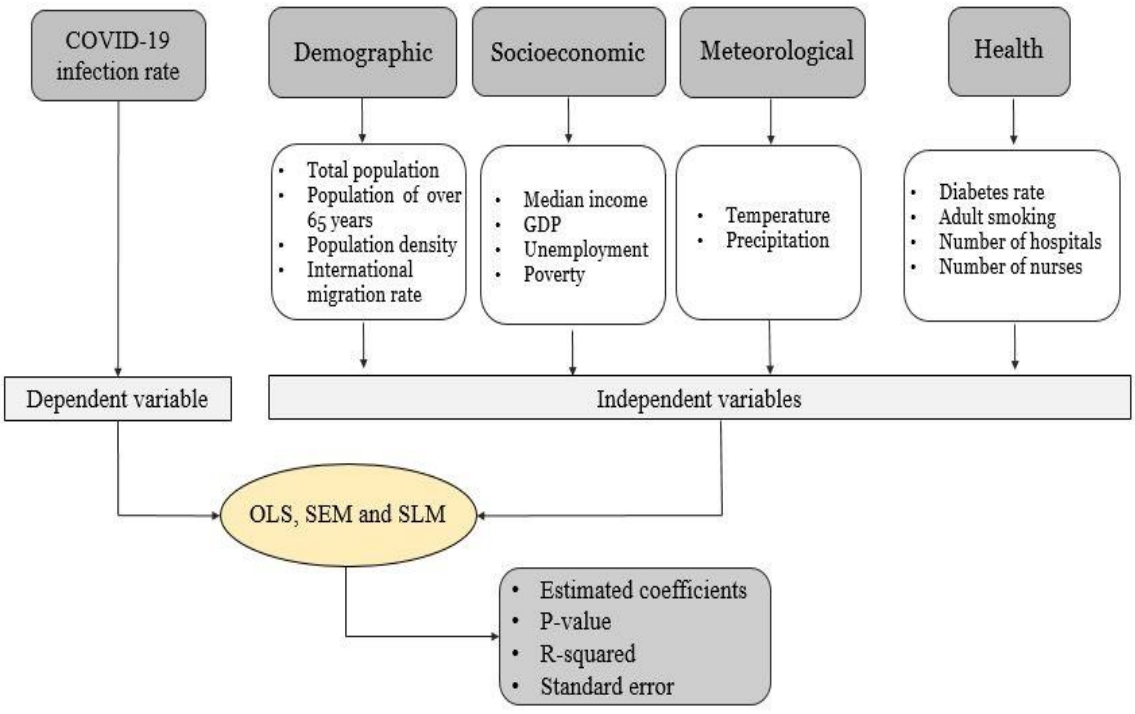

Figure 2. The general workflow for this research

3. RESULTS AND DISCUSSIONS

In this geospatial-based research, a range of various variables including demographic, socioeconomic, meteorological, and health-related factors have been considered in the modelling process as the most potentially influential factors on COVID-19 incidence rate in Canada. To fully reflect the spatial autocorrelation of the OLS regression model, the Spatial Error Model (SEM) and the Spatial Lag Model (SLM) were conducted, in which the variable of COVID-19 incidence rate is considered as the dependent variable, and fourteen parameters were applied as independent or explanatory variables.

After utilizing three spatial regression models and correlation analysing among fourteen explanatory variables, three parameters including median income, unemployment rate and diabetes were selected as the main factors contributing to COVID-19 in Canada (Table 2).

The Coefficient indicates the type and strength of relationship between explanatory and dependent variables. The asterisk (*) in P-value represents that the coefficients are statistically significant.

For OLS model, the final variables have low multi-collinearity and VIFs, and are positively related to COVID-19 incidence rate. The estimated coefficient for median income and diabetes is positive which means that the increase in household income and diabetes will cause rising in disease transmission rate in Canada. This coefficient also demonstrates that diabetes rate is the most effective factor associated with COVID-19 incidence rate, followed by median income. Moreover, a rise in unemployment rate results in decreasing the rate of disease transmission in Canada. An adjusted R^2 of 0.38 in the OLS model means that 62% of COVID-19 incidence rates are caused by uncertain factors.

The least-square regression model interpretation is based on multi-collinearity, robust probability, adjusted R, and Akaike Information Criteria (AIC). The statistically significant factors were represented by the robust probability, which indicates their significance in the model. In order to examine the VIF values and robust probability, the OLS model was run several times until all the redundant variables (among fourteen) were removed from the model. This process continued until narrowing down to the significant variables. Then, the AIC was utilized to determine the best model.

By considering the spatial dependence, the performance of OLS was significantly enhanced using SEM and SLM models. The goodness of fit statistics such as R-squared and AIC can be used for the models to estimate the fitting degree of regressions (Liu et al. 2018). The adjusted R^2 for these models increased by 0.58 and 0.61, respectively (Table 3). As can be seen, the SLM model has the best fitting impact of the three models.

Variable	Coeff.	T-statistic	P-value
Intercept	-0.0866	-2.1570	0.0190*
Median income	0.0002	2.9939	0.0005*
Unemployment	-0.6506	-2.2832	0.0385*
Diabetes	0.4670	2.6009	0.0280*

Table 2. Summary statistics of global OLS model on selected explanatory variables in modelling COVID-19 in Canada.

	OLS	SEM	SLM
Adj R^2	0.38	0.58	0.61
AICc	-63.42	-75	-73

Table 3. Comparison of the global regression models.

These global regression models are effective in spatial modelling to investigate the correlation between COVID-19 incidence rate and possible explanatory parameters based on the assumption that this relationship and correlation is a stationary spatial model (Miller, 2012; Shariati et al. 2020).

The OLS model for this study was successfully constructed with COVID-19 cases as the response variable and the risk factors including median income, unemployment rate and diabetes as independent variables. This means that other variables have not significantly correlated with the spread of COVID-19 in Canada. Although other studies indicated that some demographic and environmental factors such as population over 65 and temperature may cause severity in disease incidence rate, no significant contribution was found between these factors and COVID-19 transmission rate in Canada.

4. CONCLUSION

This study applies the global spatial modelling to determine the most influential factors affecting COVID-19 incidence rate in Canada. Three global regression models including OLS, SLM and SEM have been used and compared to examine the spatial patterns of disease transmission rates. The OLS model demonstrated a low adjusted R^2 of 0.38 compared to the SEM and SLM models, 0.58 and 0.61, respectively. This result indicates that almost 39 percent of COVID-19 incidence rate still remain unexplained or caused by unknown variables. Therefore, the global regression models might not be efficient enough for determining the spatial interactions between response and independent variables.

The OLS model considers stationarity relationship across the area in contrast to non-stationary assumption of the GWR model. In addition, while the OLS model does not adequately represent the observed spatial variations, the GWR model is able to capture local patterns and provide better overall fit (Foley et al. 2009; Fotheringham, 2009).

The global OLS model assumes that the relationship between the response variable (COVID-19) and explanatory parameters is constant across an area which means these relationships do not alter over space. Another assumption is that COVID-19 incidence rates at the subnational level are independent of each other. This model does not consider the spatial dependences, and thus, some explanatory variables that have spatial correlations may be omitted from the model.

Therefore, our next study will consider utilizing the local models such as GWR and MGWR which assume the spatial non-stationary and heterogeneity in space, and are capable of estimating the location variations of these relationships.

Moreover, in order to build a more accurate model for predicting the rate of COVID-19 in Canada, more variables should be considered. However, due to the lack of access to other data such as travel history and air pollution for all the provinces in Canada, we ignored these factors in the current study.

REFERENCES

Anselin, L. 2003. Spatial externalities, spatial multipliers and spatial econometrics. Int. Reg. Sci. Rev. 26 (2), 153–166.

Chi, G., & Zhu, J. 2008. Spatial Regression Models for Demographic Analysis. *Population Research and Policy Review*, *27*(1), 17–42. https://doi.org/10.1007/s11113-007-9051-8.

Dickin SK, Schuster-Wallace CJ, Elliott SJ. 2013. Developing a vulnerability mapping methodology: Applying the waterassociated disease index to dengue in Malaysia. PLoS One. 8(5):e63584.

ESRI. 2017. Regression analysis tutorial for ArcGIS 10. ESRI Press.

Fang, C., Liu, H., Li, G., Sun, D., & Miao, Z. 2015. Estimating the impact of urbanization on air quality in China using spatial regression models. *Sustainability (Switzerland)*, *7*(11), 15570–15592. https://doi.org/10.3390/su71115570.

Foley R, Charlton MC, Fotheringham AS. 2009. GIS in health and social care planning. In: Handbook of Theoretical and Quantitative Geography, UNIL-FGSE-Workshop series (2). Lausanne: Univ.de Lausanne- Faculté des géosciences et de l'environnement.. 73-115. http://mural.maynoothuniversity.ie/2990/1/FoleyCharltonFotheringham_Final.pdf

Fotheringham AS. Geographically weighted regression. 2009. In: The SAGE Hand Book o Spatial Analysis. London: Sage Publications.

Hassan, M.S., Hossain, M.A., Tareq, B. F., Bodrud-Doza, M., Tanu, S.M., Rabbani, K.A., 2021. Relationship between COVID-19 infection rates and air pollution, geo-meteorological, and social parameters. Environ. Monit. Assess. 193 (29).

Hutcheson GD. 2011. Ordinary Least-Squares Regression. In: Moutinho, L. and Hutcheson, G.D., The SAGE Dictionary of Quantitative Management Research, SAGE Publications, Thousand Oaks. 224-8.

Izón, G.M.; Hand, M.S.; Mccollum, D.W.; Thacher, J.A.; Berrens, R.P. 2016. Proximity to Natural Amenities: A Seemingly Unrelated Hedonic Regression Model with Spatial Durbin and Spatial Error Processes. Grow. Chang. 47, 461–480.

Jain, S., Sannigrahi, S., Sen, S., Bhatt, S., Chakraborti, S., & Rahmat, S. 2019. Urban heat island intensity and its mitigation strategies in the fast-growing urban area. *Journal of Urban Management*, 5856. https://doi.org/10.1016/j.jum.2019.09.004.

Liu, R., Yu, C., Liu, C., Jiang, J., and Xu, J., 2018. Impacts of Haze on Housing Prices: An Empirical Analysis Based on Data from Chengdu (China). Int. J. Environ. Res. Public Health, 15, 1161. doi: https://doi.org/10.3390/ijerph15061161.

Miller JA. Species distribution models Spatial autocorrelation and non-stationarity. Progress in Physical Geography. 2012; 36:681-92.

Mollalo, A., Vahedi, B., Rivera, K.M., 2020. GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. Sci. the Total Envir. 728 (2020) 138884.

Neill, D.B. 2006. Detection of spatial and spatio-temporal clusters. Tech. Rep. CMU-CS-06-142. [Ph.D. Thesis]. Pittsburgh, PA: Carnegie Mellon University.

Oshan, T.M., Li, Z., Kang, W., Wolf, L.J., Fotheringham, A.S., 2019. Mgwr: a Python implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale. ISPRS Int. J. Geo Inf. 8 (6), 269.

Pickle LW. Spatial Analysis of Disease. 2002. In: Biostatistical Applications in Cancer Research. vol. 113. Boston, MA: Springer.

Rahman, M.H., Zafri, N.M., Ashik, F.R., Waliullah, M., Khan, A., 2021. Identification of risk factors contributing to COVID-19 incidence rates in Bangladesh: A GIS-based spatial modeling approach. Heliyon 7(2021) e06260.

Rex, F.E., Borges, C.A.S., Kafer, P.S., 2020. Spatial analysis of the COVID-19 distribution pattern in Sao Paulo State, Brazil. Ciência & Saúde Coletiva, 25(9):3377-3384.
Sannigrahi, S., Pilla, F., Basu, B., Basu, A.S., Molter, A., 2020. Examining the association between socio-demographic composition and COVID-19 fatalities in the European region using spatial regression approach. Sustainable Cities and Society 62(2020) 102418.

Shariati M, Jahangiri-rad M, Mahmud Muhammad F, Shariati J. 2020. Spatial Analysis of COVID-19 and Exploration of Its Environmental and Socio-Demographic Risk Factors Using Spatial Statistical Methods: A Case Study of Iran. Health in

Emergencies and Disasters Quarterly. 5(3): 145-154. http://dx.doi.org/10.32598/hdq.5.3.358.1

Wang, Y.; Wang, S.J.; Li, G.D.; Zhang, H.G.; Jin, L.X.; Su, Y.X.; Wu, K.M. 2017. Identifying the determinants of housing prices in China using spatial regression and the geographical detector technique. Appl. Geogr. 79, 26–36.

Ward, M.D., Gleditsch, K.S. 2018. Spatial regression models. 155. Sage Publications.

World Health Organization (WHO). 2020. Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19). Retrieved from. https://www.who.int/docs/default source/coronaviruse/who-china-joint-mission-on covid-19-final-report.pdf.

WHO. Director-General's opening remarks at the media briefing on COVID-19 [Internet]. 2019 [Updated 2020 March 11]. Available from: https://www.who.int/director-general/ speeches/detail/who-director-general-s-opening-remarksat-the-media-briefing-on-covid-19---11-march-2020.

WHO. Coronavirus disease (COVID-19) pandemic [Internet]. 2020 [Updated 2020 November 23]. Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

Wu, X., Nethery, R.C., Sabath, B.M., Braun, D., Dominici, F. 2020. Exposure to Air Pollution and COVID-19 Mortality in the United States. https://doi.org/10.1101/2020.04.05.20054502.

Zhang, C.H., Schwartz, G.G., 2020. Spatial Disparities in Coronavirus Incidence and Mortality in the United States: An Ecological Analysis as of May 2020. J. Rural Heal. 36, 433–445. https://doi.org/10.1111/jrh.12476.

Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y., Song, B., Gu, X., Guan, L., Wei, Y., Li, H., Wu, X., Xu, J., Tu, S., Zhang, Y., Chen, H., & Cao, B. 2020. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet*, *395*(10229), 1054–1062.