Colorectal cancer risk factors in north-eastern Iran: A retrospective cross-sectional study based on geographical information systems, spatial autocorrelation and regression analysis

Ladan Goshayeshi,1 Ali Pourahmadi,2 Majid Ghayour-Mobarhan,2 Soheil Hashtarkhani,4,5 Sajad Karimian,6 Reza Shahhosein Dastjerdi,1 Babak Eghbali,7 Efat Seyfi,7 Behzad Kiani5

1Department of Gastroenterology and Hepatology, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad; 2Department of Computer Engineering, University of Science and Arts of Yazd, Yazd; 3Metabolic Syndrome Research Center, Mashhad University of Medical Sciences, Mashhad; 4Student Research Committee, Mashhad University of Medical Sciences, Mashhad; 5Department of Medical Informatics, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad; 6Student Research Committee, Urmia University of Medical Sciences, Urmia; 7Department of Health, Mashhad University of Medical Sciences, Mashhad, Iran

Abstract

Colorectal cancer (CRC) is the second most common cancer among females and the third most common malignancy in males in the world. No single risk factor has been identified, but there are many interrelated factors that together cause the disease. This retrospective, cross-sectional study aimed to identify potential spatial factors contributing to its geographical distribution. Data concerning 1,089 individuals with CRC from the Khorasan-Razavi Province in Iran, located in the North-East of the country, were obtained from the national CRC registry. Local Moran’s I statistic was performed to identify clustered areas of CRC occurrence and ordinary least squared regression was calculated to predict frequency of the disease based on a set of variables, such as diet, body mass index (BMI) and the proportion of the population ≥50 years of age. Some dissimilarities related to the geography in the province and also some neighbourhood-related similarities and dissimilarities of CRC occurrence in the city of Mashhad were found. A significant regression equation was found (F(4,137)=38.304, P<.000) with an adjusted R2 of 0.6141. The predicted CRC frequency was -58.3581 with the coefficients for average BMI=+1.594878; fibre intake=-0.610335; consumption of red meat +0.078970; and ≥50-year age group =+0.000744. All associations were statistically significant, except the consumption of red meat one. The study results illuminate the potential underlying risk factors in areas where the CRC risk is comparatively high and how the CRC risk factors may play a role in CRC geographic disparity. Further research is required to explain the patterns observed. We conclude that people should include more fibre in their daily diet and decline their BMI to decrease risk of CRC.

Introduction

Cancer is a major public health problem and the second most common reason for global mortality (Siegel et al., 2019). Colorectal cancer (CRC) is the second most common cancer among females and the third most common malignancy in males worldwide (Fitzmaurice et al., 2017). CRC develops when the DNA mutations occur in cells of the colon or rectum making cells unable to control their growth and division. In many cases, these mutated cells die without known cause or following attack by the immune system. However, sometimes the mutated cells escape
immune surveillance and start to grow out of control, they form tumour(s) in the colorectal part of the intestines, often resulting in metastatic growth also in other organs (Janssen et al., 2017). The disease represented the loss of 15,800,000 disability-adjusted life years in 2013, 56% of which in middle- and low-income countries and 44% in industrialized countries (Fitzmaurice et al., 2015). There is a high probability that these figures are skewed in the direction of a higher percentage in low-income countries.

In Iran, CRC is the fourth most malignant growth in men and the third most common cancer disease in women (Mansori et al., 2018). No single hazardous factor has been identified, but individual factors (e.g., sex, age and family history), way of life (e.g., diet, physical activity and smoking), capability of medical services (e.g., financial aspects and access to medical services) and environmental factors all include CRC determinants (DeChello and Sheehan, 2007; Haggar and Boushey, 2009; Brenner et al., 2014; Shah et al., 2014; Adeola et al., 2018). Many factors are interrelated and environmental elements may strongly exacerbate the impact of other hazardous factors. Indeed, patients regularly exposed to environmental variables (i.e. air contamination, pesticides, water pollution and heavy metals) might be at higher risk for CRC (Chang et al., 2015; Yuan et al., 2016; Nawi et al., 2019). Therefore, spatial examination of the CRC incidence is of great significance in the study of the epidemiology of the disease.

Figure 1. The study area in North-East Iran: Khorasan-Razavi Province.
Results, when substantiated, would be useful as guide to policy makers for more efficient interventions aimed at disease prevention. Geographical Information System (GIS) empowers investigations on the spatial distribution of diseases (Kiani et al., 2017b). Previous studies have utilized several methods based on GIS for spatial examination of CRC occurrence. Shah et al. (2014) carried out a spatial analysis of CRC cases in Kuala Lumpur, Malaysia covering the years 1995-2011 utilizing spatial autocorrelation methods. Their investigation demonstrated Chinese as having the highest number of registered cases and that they were concentrated in northern Petaling and Kuala Lumpur districts, an important distinction in the spatial dissemination of CRC occurrence. Spatial autocorrelation methods have been used to investigate whether the observed value of a variable in one territory is independent of values of the variable in neighbouring areas (Griffith and Paelinck, 2018). A study by Halimi et al. (2019) highlighted clustered regions of CRC incidence based on a Local Moran’s I in Iran’s Hamadan Province from 2007 to 2014. Applying Anselin Local Moran’s I to a large number of colorectal surgical procedures conducted in New South Wales, Australia covering 2002 to 2013, Bagheri et al. (2017) found hotspots and coldspots of this disease of various intensity in outer, regional and remote areas of New South Wales, Australia.

Performing spatial cluster analysis based on various statistical approaches, Al-Ahmadi et al. (2013) and Li et al. (2017) working in southern China and Saudi Arabia, respectively, found widespread clusters indicating a link between geography and cancer occurrence, while Torres et al. (2018) used ordinary least squares (OLS) regression to explore associated environmental and socio-economic risk factors of CRC in the city of Baltimore, USA from 2000 to 2010. OLS regression analysis lends itself to modelling linear relationships leading to a better understanding of the factors behind observed global spatial patterns and outcome prediction (Lawson et al., 2016). To our best knowledge, spatial variation of CRC incidence has not been investigated in Khorasan-Razavi Province, Iran. Furthermore, we did not find a study applying OLS regression for associations between CRC and nutritional and lifestyle variables. Therefore, we undertook a study examining the spatial variations in the pattern of CRC occurrence in this province.

Figure 2. Spatial autocorrelation of ordinary least squared residuals.
Materials and Methods

Study area

This study was conducted in Khorasan-Razavi Province including Mashhad, which is the capital of the province. Khorasan-Razavi is one of the 31 provinces of Iran. It covers an area of 118,854 km² and the population of the province according to the latest National Census was 6,434,501 people in 2015, 3,372,660 of whom lived in the provincial capital (Statistical Center of Iran, 2018). Figure 1 shows the location of the study area.

Data sources

Data were obtained from three different sources with three different spatial scales (individual, census tract and county). Individual CRC data were obtained from the population-based cancer registry in Khorasan-Razavi Province and patient addresses, gender and ages were included in the dataset. There were 1,089 cases of diagnosed CRCs in the study area between 2016 and 2017. The neighbourhood division and population of them were obtained from the City Council in Mashhad City. Although the data were provided at the census block level, they were combined to the neighbourhood level for better representation and inference. Data regarding clinical factors, body mass index (BMI) and dietary habits, including daily consumption of red meat and fibres, were obtained from the MASHHAD cohort study (Ghayour-Mobarhan et al., 2015), which started in 2010 and will continue until 2020.

Study subjects

Mashad cohort study participants were drawn from three regions in Mashhad City, using stratified cluster random sampling (Ghayour-Mobarhan et al., 2015). Each region was divided into nine sites using the divisions of Mashhad Healthcare Center. The total dataset included 9,761 individuals from the MASHHAD cohort (Ghayour-Mobarhan et al., 2015). County divisions and their populations at the province level were obtained from the Statistical Center of Iran (2018). We geocoded all addresses manually using Google’s MyMaps software (https://www.google.com/maps/about/mymaps) preserving privacy by using a geomasked technique. The exact address of each patient was only used for the spatial analysis, while the accuracy was less than 1 km when mapping the areas, as a radius of 0.5 km around the homes where then used.

Figure 3. Geographical distribution of the colorectal cancer incidence rate at county level in Khorasan-Razavi Province, Iran.
Statistical analysis

We used Local Moran’s I, Kriging, and OLS in this study. The CRC incidence rates were calculated using the total population and the number of cases in each county of the province and part of Mashhad City. There are 69 counties and 142 urban neighbourhoods in the study area. Local Moran’s I statistic was performed to quantify spatial autocorrelation of CRC frequency at the county and neighbourhood levels. Z-scores and P-values were determined to know whether apparent similarities (spatial clusters of either high or low values) or dissimilarities (shown as spatial outliers) would be more pronounced than expected for random distribution. With a null hypothesis of the CRC cases randomly distributed across the study area, High-High (HH) and Low-Low (LL) regions would indicate that the study area had similar high or low rates of CRC, while High-Low (HL) and Low-High (LH) regions would imply regions with dissimilar rates as how in New South Wales, Australia by Bagheri et al. (2017). In other words, the HH and LL areas indicate the presence of CRC clusters, while HL and LH areas indicate outliers.

Kriging is an interpolation algorithm that creates a raster surface based on known point values. We spatially interpolated the levels of BMI, red meat consumption and fibre intake by Kriging (Kiani et al., 2018) with 500 nearby points. Then we converted the results into vector polygon layers by creating a point layer of pixels and then spatially joined the points to a layer of Mashhad City neighbourhoods using ArcGIS (ESRI, Redlands, CA, USA). The average of each variable (level of BMI, red meat consumption, fibre intake) were computed for each polygon.

A regression model considering CRC frequency as the dependent variable was created based on an OLS approach. The independent variables consisted of the proportion of the population ≥50 years of age, average BMI, average daily red meat consumption, and average of daily fibre intake. Global Moran’s I was used to find out whether the OLS residuals were autocorrelated, which would tell us that some important variables had been missed. However, in this study, they were not significantly different from the random distribution (Figure 2). The coefficient of determination (R²) varies between 0 and 1 and shows the percentage variation in the dependent variable and is therefore used to establish the performance of regression models. The problem is that R² would either stay the same or increase with addition of variables, even if they would not have any relationship with the output variable. However, we used the adjusted R² in this study to find out if the model is penalized when variables are added which do not improve the model. This is an indicator that provides a measure of the goodness of fit (Harel, 2009; Miles, 2014).

We also calculated the variance inflation factor (VIF), which detects multicollinearity which signifies a correlation between the predictors (i.e. the independent variables) in regression analysis. A high VIF value (near 10 or above) is not acceptable, while values around 1 indicates that the regression is reliable and working well.
ArcGIS 10.6 was used for all spatial analyses and Microsoft Excel 2013 for the descriptive ones. The layers were projected using the WGS_1984_UTM_Zone_40N system, and the significance level at $\alpha=.05$ was set for all calculations.

Results

A total of 1,089 CRC cases with an average age of 60.3 years were recorded in Khorasan-Razavi Province during 2016-2017. The incidence rate was 16.92 per 100,000 people for the province and 20.93 per 100,000 people for Mashhad City. Men showed a higher CRC rate (57%) compared to women (43%). The geographical distribution of the disease in the study area is shown in Figure 3, where Bakhezr County stands out with the highest incidence rate of CRC.

The spatial cluster analysis of CRC cases at the county level in Khorasan-Razavi Province revealed three significantly HL outlier areas in Sabzevar, Fariman, and Gonabad areas. Further, there were four significant LH outliers in Baba Velayat, Salami, Mian Velayat, and Taybad (Figure 4). Figure 4 also shows that there was no clustering at all of CRC in Khorasan-Razavi Province as a whole. However, at the neighbourhood level in Mashhad City we found three significant HH clusters (hotspots) of CRC incidence in Rezashahr, Sarafrazan and Nohe-Leoshato. Further, there were also significant LL (coldspots) and they were seen in south-eastern Mashhad City (Figure 5). Figure 5 also shows some outlier regions (HL and LH) of CRC incidence in Mashhad City.

Figure 6 shows the spatial distribution of independent variables used to build the OLS model for CRC frequency as the response variable. Table 1 shows the result of performing OLS model to predict CRC frequency at the neighbourhood level in Mashhad City. Table 1 clarifies that the coefficients regarding the proportion of the population ≥50 years of age, the average BMI and the daily average of fibre intake were significantly associated with CRC frequency in Mashhad City. However, Table 1 emphasizes that the level of average daily red meat consumption among the Mashhad population was not associated with CRC at a statistically significant level. The adjusted $R^2$ of the OLS model was 0.614157, which indicates that more important variables affecting CRC occurrence may exist. However, the VIF values calculated were all low (<2) underlining that the independent variables were not correlated.
Discussion

To the best of our knowledge, this is the first research identifying a spatial pattern of CRC incidence in Khorasan-Razavi Province, Iran. The analysis at the provincial level found a significant spatial variation of CRC incidence with some HL and LH outliers as seen in Figure 4. Furthermore, we identified several CRC clusters in Mashhad City as seen in Figure 5. However, to be a statistically significant hotspot, a HH area should be surrounded by areas characterized by equally high values. The spatial outliers seen in Figure 4 included both HL and LH areas and these clusters and outliers should be investigated to know why their CRC occurrences differ. As the results of this study (Table 1) confirms, dietary factors might be a reason. However, previous research reports show other possible reasons, such as poor access to health care services (Ambroggi et al., 2015; Kiani et al., 2017a), hereditary reasons (Goshayeshi et al., 2017, 2018), environmental factors (Jacquez and Greiling, 2003) and socioeconomic inequalities (Goungounga et al., 2016). For example, patients with poor access to health care services may have polyps that can cause CRC due to delay or lack of treatment. Inheritance is another risk factor for developing CRC, so areas that are currently at higher risk might also have had a large number of CRCs in the past. In addition, people with poor socio-economic status might have poor dietary and life style which could develop CRC.

The OLS approach used here confirms some of the assumptions about various associations with CRC, such as BMI, average daily fibre intake, and high age (≥50) (Table 1), with obesity and overweight having the greatest impact with CRC in this study. Wang et al. (2019) have studied this issue in a large number (0.5 millions) of men and women in China showing that high BMI is associated with an increased risk for CRC. A high BMI score is

![Visual representation of the independent variables used in ordinary least squared regression analysis.](image-url)
also associated with premature general mortality (Shaukat et al., 2017). The present study recommends that weight-loss programmes should be implemented as a top priority to reduce the incidence and mortality of CRC. Table 1 shows that the amount of daily fibre intake was inversely related to the incidence of CRC. Previous studies have confirmed this inverse relationship (Park et al., 2005; Kunzmann et al., 2015). There was a weak significant relationship between the population ≥50 years of age and CRC incidence (Table 1). Although the CRC incidence trend has been declining over the past few decades in the older population, incidence rates for adolescents and young adults have steadily increased. Several factors may well explain this apparent epidemic in the young, namely a lack of routine screening and emerging lifestyle issues such as obesity, lack of exercise and particular diets (Connell et al., 2017). Our study showed that there was no significant association between CRC occurrence and red meat consumption. Contrary to our study, a review article recently reported that this association has been proven by some authors (Sasso and Latella, 2018). Our results may not be significant because we did not consider processed meats. Alexander et al. (2015) conducted a meta-analysis of the effect of red meat consumption and CRC showing that there was no significant relationship when only red meat was considered. However, when the consumption of processed meat and natural red meat were accounted for together, this effect was significant (Alexander et al., 2015). There are also some other recent systematic reviews and meta analyses which support our findings about red meat consumption and CRC (Han et al., 2019; Vernooij et al., 2019; Zeraatkar et al., 2019).

The adjusted R² for the OLS model was 0.614157 in this study. This figure indicates that there may exist more effective variables affecting CRC, so far not identified. Looking for possible associations, trace elements (Nawi et al., 2019) and heavy metals, such as arsenic, cadmium, chromium and nickel (Yuan et al., 2016), have been mentioned. One noted source of environmental exposure to heavy metal could be via the human food chain through contact with contaminated soil and water (Chen et al., 2015), and recent studies in North-East Iran have showed moderate, sometimes very high, heavy-metal contamination of the agricultural soil, as well as elevation of lead in water in some areas of Mashhad City (Peiravi et al., 2013). Other possible risk factors are alcohol use and smoking (Fagunwa et al., 2017). In addition, genetic factors play a significant role for the incidence of CRC (Goshayeshi et al., 2017). Finally, the socioeconomic status differs geographically and may have an impact on the epidemiology of cancer incidence as indicated by Mao et al. (2001) and Goungounga et al. (2016). We suggest that in addition to the independent variables considered in this study, future researchers also investigate these additional variables when data are available.

This study has some limitations. The precision of cancer registration in Iran is not high due to incomplete electronic registries, which may have caused some omissions of CRC patients in our study. Second, there may be patients who have moved to other provinces for treatment and therefore are not included in our dataset that covers only Khorasan-Razavi Province. However, since Mashhad City is the most advanced urban area in the eastern part of Iran for cancer treatment, this is very unlikely. Finally, because the ArcGIS geocoding tool could not geocode Persian addresses, we were forced to geocode them in other ways. One of the authors geocoded the addresses by Google’s MyMaps software, while another author checked the addresses to avoid error; in case of disagreement the address registration was checked by a third person.

Table 1. Results of ordinary least squared regression analysis.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>P-value</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-58.353158</td>
<td>0.006621*</td>
<td>-</td>
</tr>
<tr>
<td>Number of people aged ≥50</td>
<td>0.000744</td>
<td>0.000000*</td>
<td>1.143303</td>
</tr>
<tr>
<td>Average BMI (kg/m²)</td>
<td>1.594878</td>
<td>0.028375*</td>
<td>1.303889</td>
</tr>
<tr>
<td>Daily consumption of red meat (g)</td>
<td>0.078970</td>
<td>0.091707</td>
<td>1.695972</td>
</tr>
<tr>
<td>Daily consumption of fibre (g)</td>
<td>-0.610335</td>
<td>0.006597*</td>
<td>1.591735</td>
</tr>
</tbody>
</table>

VIF, variance inflation factor; BMI, body mass index. *Indication of statistical significance. The adjusted R² was 0.614157.

Conclusions

The study examined the potential relationship between CRC frequency with individual factors including BMI, dietary patterns, and old age. The regression analysis showed that these factors could be associated with CRC. To improve the situation, we recommend people to focus on intake of fibre in their daily diet and decrease their BMI. The study illuminates these factors as potential underlying risk factors in areas where the CRC risk is comparatively high and investigates which risk factors may play a role in CRC geographic disparity.

References

DeChello LM, Sheehan TJ, 2007. Spatial analysis of colorectal...


