Original Article
Middle East Journal of Cancer; January 2020; 11(1): 91-98

Pattern of Gastric Cancer Incidence in Iran is Changed after Correcting the Misclassification Error

Mohamad Amin Pourhoseingholi*, Hadis Najafimehr♦, Nastaran Hajizadeh*, Mohammad Reza Zali***

**Gastroenterology and Liver Diseases Research Center, Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran*
***Basic and Molecular Epidemiology of Gastrointestinal Disorders Research Center, Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran*

Abstract

Background: The main aim of the present study was to map the real high risk regions of gastric cancer (GC) using corrected data for misclassification error in all Iranian provinces.

Method: In this cross-sectional study, the data were extracted from the reports of Ministry of Health and Medical Education and previous studies on correcting GC registered data including 30 provinces in 2008. The information about socioeconomic factors was extracted from the statistical centers of Iran. To estimate the model parameters, the Bayesian approach was used with regards to spatial correlation due to adjacent effects.

Results: The southern and northern provinces were introduced as high-risk regions and the central provinces were introduced as low-risk regions. The mean household income was inversely associated with the risk of GC.

Conclusion: The real high-risk regions of GC in Iran are the north and south border provinces which should be considered by health policy makers. It is also necessary to correct misclassified registered data which can lead to seduction in health service allocation.

Keywords: Misclassification correction, Disease mapping, Risk factor, Gastric cancer, Iran

♦Corresponding Author:
Hadis Najafimehr, MSc
Research Institute for Gastroenterology and Liver Diseases
Aerabi Ave., Tabnak Str., Evin, Tehran, Iran
Postal code/ P.O. Box: 1985714711
Tel: +98-21-22432525
Fax: +98-21-22432517
Email: ghfbb.journal@gmail.com

Introduction

Various parts of the world are faced with the high incidence of Gastric cancer (GC). As a general comparison, GC incidence rate in Eastern Europe and Eastern Asia is higher than North America and Africa.[1] Additionally, the incidence rate is higher in males than females.[2] In spite of the various studies on GC, our knowledge on GC prevention is yet to be completed.[3]

In regional epidemiologic studies,

Received: January 02, 2019; Accepted: September 11, 2019

disease mapping is usually employed to investigate the incidence and prevalence of diseases. This approach is important because it can a) describe the variation of geographical disease risk, b) introduce the existent regional risk factors, c) reveal the unequal access to health services, and d) identify the high-risk regions.[4] Via disease mapping, the disease distribution and spread may become apparent in each geographical region and the cluster of high risk points becomes distinguishable. The clusters by different risk distinguish, is due to the presence of significant difference between disease risk ranges in different geographical points.[5] Therefore, the resultant maps are very useful in the surveillance and monitoring of the disease.

It is clear that the health outcomes in a special location are similar. The outcomes that occur under the same environmental circumstances and lifestyles are even more similar. This similarity, which explains the presence of spatial correlation, may sometimes not conformed, probably due to the existence of misclassification error in the registered data.[6,7] Misclassification is the difference between the registered data and the actual amount. This problem is created when the data belonging to a special point is registered at an adjacent rather than the actual point.[8,9] Misclassification error will lead to unreliable and inaccurate estimations and also may lead to underestimation of disease risk and finally will lead to an inaccurate and incorrect allocation of health budgets. In Iran, as a developing country, this problem exists but there is a progressive trend towards its correction.[10] In this regard, there exist certain statistical solutions that can correct the registered data.[11-13]

Having the accurate primary data is important for creating the maps and if the data are misclassified, the notice understood by disease mapping may become misleading. Due to misclassification error in cancer registered data in Iran, the previous attempts at identifying GC high risk regions have been inconsistent and incomplete. To address this gap, the present study was designed to i) introduce real Iranian GC high risk regions , ii) compare the relative risk (RR) of GC using corrected and uncorrected data for misclassification error, and iii) evaluate the socioeconomic risk factors of GC.

Materials and Methods

Data collection

In Iran, since 1984, the data for cancer incidence using hospital records and diagnosis have been registered according to the International Classification of Diseases by Ministry of Health and Medical Education (MHME). All registered GC cases were collected from MHME according to ICD-10: C16 diagnostic code for 30 provinces in 2008.[14] The misclassification in the Iranian cancer registered data has already been reported and there exist certain approaches to reduce the problem. To follow the tread on GC data, the misclassification rate in the registered data between each two neighboring provinces was estimated and the incidence of GC for each province was then re-estimated with the aid of Bayesian method. Therefore, the corrected GC data were introduced in all 30 provinces.[15] In the present cross sectional study, the authors used the corrected data in addition to all previous registered cases.

The population at risk and the information regarding the mean household income (MHI) and unemployment rate (UER) were obtained from the census data gathered by statistical center of Iran (SCI) in 2006. The census of population and housing is performed every 5 years so as to prepare the socioeconomic data and various related surveys.

The study protocol was approved by the institutional review board of Gastroenterology and Liver Diseases Research Center, Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran.

Statistical analysis

The model of misclassification data (before misclassification correction) and the model of corrected misclassification data (after misclassification correction) were compared. To estimate the disease incidence rate, smoothed standardized

Table 1. Summary of gastric cancer incidence according to misclassification correction status in Iran (2008)

		Total (%)	Min	Max	Median	Mean	SD
Before misclassification correction	Women	2243 (30.31)	9	567	44	74.77	104.25
	Men	5156 (69.68)	17	1131	103	171.87	218.70
	Both genders	7399 (100)	26	1698	150.50	246.63	321.97
After misclassification correction	Women	2243 (30.32)	19	511	55	74.77	88.11
	Men	5154 (69.67)	45	1019	118	171.80	179.58
	Both genders	7397 (100)	67	1530	178	246.57	266.74

incidence rate (SIR) is usually used.[16] Accurate estimate of the SIR can be obtained by obtaining information from adjacent regions and with the aid of hierarchical and fully Bayesian method.[17] In case of low disease incidence in the under-study span, it is assumed that o_i is the number of observed cases in area and i follows Poisson (E_iR_i) distribution, where E_i are the expected cases and RR_i is the relative risk (RR). The used random effect model is as follows:

$$\log(RR_i)=\log(E_i)+\alpha+\beta_i+\gamma_i+\theta^T x_i$$

where α is an overall mean, β_i is the correlated heterogeneity term, γ_i is the uncorrelated heterogeneity term, x_i is the covariates vector, and θ is the regression coefficients vector.[4,18] The model was fitted using Markov Chain Monte Carlo (MCMC) simulation methods with conditional autoregressive (CAR) and Normal priors distribution for spatial and other non-spatial terms, respectively.[19,20] To compare the RRs between the two models of misclassification correction, the paired t-test was used. After gaining convergence, the posterior estimates of model parameters were obtained. The results were achieved by the WinBUGS software, and Arc GIS software was used to plot the maps.[21] The level of 0.05 was considered as the significant level.

Results

The results were calculated according to the data misclassification correction status (before and after misclassification correction). Prior to correction, there were 7399 GC cases, while after the correction, there were 7397 cases. In table 1, certain descriptive statistics of incidence are listed according to gender. The expected GC cases by province were calculated with the aid of internal standardized method in table 2. Given the observed and expected numbers in tables 1 and 2, a clear difference was recognized between each correction status and gender. Tehran province (the capital of Iran) had the highest number of observed and expected incidence cases among all provinces.

Regarding women before correction, the maximum RR (1.99) belonged to Semnan and the minimum belonged to Sistan province. However, after correction, Hormozgan province had maximum RR (2.47) and Fars province had minimum (0.39). Concerning the male population prior to correction, the maximum RR (1.93) belonged to Ardebil and the minimum (0.21) was observed in Sistan, while following after correction, Ardebil province had the maximum RR (2.64) and Fars province had the minimum (0.34).

Before correction for both genders, the maximum RR (1.96) belonged to Semnan and the minimum (0.18) belonged to Sistan. Following correction; however, Ardebil had the most (2.48) and Fars had the least (0.35) RR.

The risk of GC after correction was more than before the correction regarding women (P=0.032), men (P=0.007), and both genders (P=0.035).

The distribution patterns of RR are shown in figure 1, where before correction, the northern parts of Iran were considered as high risk regions for both genders. The range of RR for men was more than women, yet the expansion of high risk regions in women was more than men. Following correction, some southern provinces were considered as high risk regions in addition to the northern parts. Western and Eastern province were no longer considered as high risk; however, prior to correction, these parts were known as high risk areas. The RRs were higher in men, while the expansion of high-risk provinces was still higher in females.

The posterior estimates of MHI (mean±SD:60.46±12.04 per 106 Rials) and UER (mean±SD:11.64±3.18) covariates are summarized in table 3, where MHI is inversely associated with GC risk in most cases, while UER had no

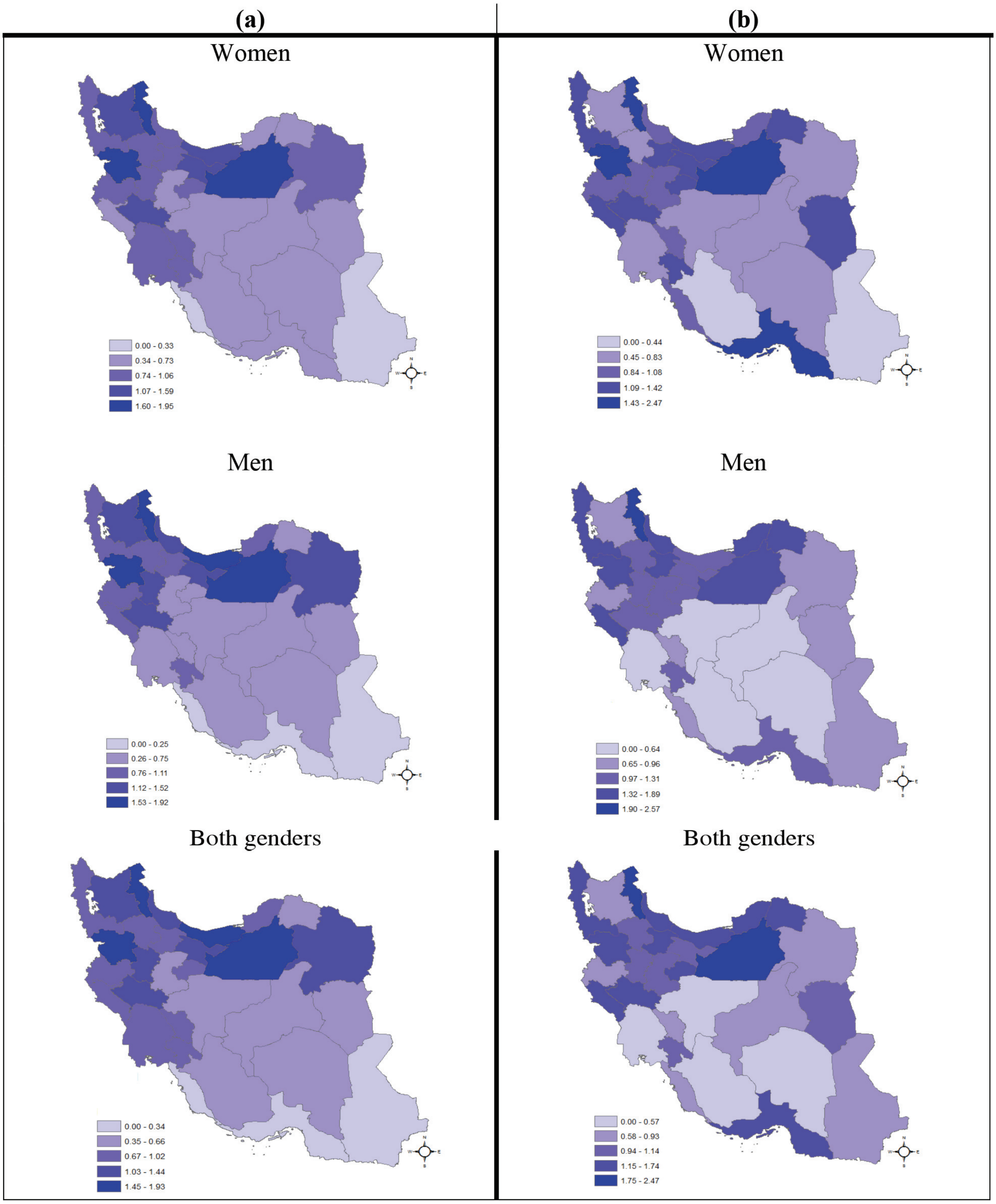

Figure 1. Spatial distribution of gastric cancer relative risk in Iran before (a) and after (b) misclassification error correction in 2008.

Table 2. Number of observed and expected gastric cancer incidences according to misclassification correction status in Iranian provinces in 2008

Province	Female				Male			
	Before correction		After correction		Before correction		After correction	
	Observed	Expected	Observed	Expected	Observed	Expected	Observed	Expected
E-Azerbaijan	148	114.89	88	114.89	339	263.68	237	263.58
W-Azerbaijan	80	92.80	122	92.80	191	211.78	291	211.70
Ardabil	69	39.83	92	39.83	172	89.29	229	89.25
Esfahan	86	143.87	77	143.87	211	333.44	188	333.31
Ilam	13	17.42	22	17.42	37	40.04	61	40.02
Bushehr	9	27.46	28	27.46	17	67.66	54	67.64
Tehran	567	421.13	511	421.13	1131	983.57	1019	983.19
Chaharmahal	24	27.97	29	27.97	46	62.78	56	62.76
S-Khorasan	13	20.58	27	20.58	22	46.74	45	46.72
R-Khorasan	194	182.76	131	182.76	550	408.36	371	408.20
N-Khorasan	15	26.96	34	26.96	42	58.60	94	58.57
Khuzestan	114	137.75	85	137.75	233	318.23	173	318.11
Zanjan	26	31.44	26	31.44	77	69.48	77	69.45
Semnan	36	18.56	36	18.56	82	43.11	82	43.09
Sistan	9	79.98	35	79.98	38	182.48	147	182.41
Fars	102	138.76	54	138.76	199	316.70	105	316.57
Qazvin	33	36.30	49	36.30	80	83.85	119	83.82
Qom	26	33.27	44	33.27	58	77.29	93	77.26
Kurdistan	85	46.34	85	46.34	180	105.19	180	105.15
Kerman	60	85.29	60	85.29	102	196.43	102	196.36
Kermanshah	57	60.03	57	60.03	117	125.56	117	125.51
Kohgiluyeh	18	20.64	24	20.64	38	46.74	50	46.72
Golestan	35	53.41	58	53.41	104	117.05	173	117.01
Gilan	81	77.52	81	77.52	245	170.49	245	170.42
Lorestan	78	55.17	78	55.17	158	126.54	158	126.49
Mazandaran	150	94.00	122	94.00	338	209.69	274	209.61
Markazi	27	43.18	44	43.18	63	97.52	102	97.48
Hormozgan	17	29.67	73	29.67	26	105.75	112	105.71
Hamadan	52	54.98	52	54.98	152	123.33	152	123.28
Yazd	19	31.06	19	31.06	48	74.64	48	74.61

significant associations except for only one case.

Discussion

Since GC is a major global disease, determining high risk regions is very important regarding preventive programs. To identify high risk regions, RR is determined in the framework of disease mapping approach. However, an important point in RR estimation is the presence of adjacent point effects. When the misclassification problem is created, this error can distort the effect of adjacent points. To show the consequences of misclassification problem in introducing high-risk areas, the current study compared two models of before and after misclassification regarding GC registered data.

Over the recent years, some studies have been carried out in Iran to solve the misclassification error. Pourhoseingholi et al. indicated that in the death data reports for GC, the percentage of undercounting was 30-40 in addition to the increase in the incidence cases.[22,23] Hajizadeh et al. observed a 34% misclassification in the registered data while reducing the misclassification error in the GC incidence data registered in Khorasan (An eastern province of Iran).[24] In another study by Hajizadeh et al. on GC registered data, the misclassification rate was calculated for all Iranian provinces. This study showed how the data for each province is mistakenly registered in other adjacent provinces. For example, they estimated that there was 72% misclassification in Sistan province (Located in the southeast) registered in Razavi Khorasan province (Located in the northeast). They inferred that Hormozgan province had the maximum percentage of changes (331.58%) after correction and there was a clear change in other southern provinces.[24] They corrected the age standardized incidence rate (ASR) data for Iranian provinces. Based on their results, regarding Hormozgan (Southern province), after misclassification correction, the ASR was estimated 20.[33] for males and 17.35 for females. However, these estimates were 4.71 and 4.02 for males and females prior to correction, respectively. These changes were also great in most of other southern provinces. As a general comparison, the pattern of ASR before and after correction is prepared in figure 2. The mentioned corrected data formed a part of the present study data and the above descriptions may explain the difference in our maps compared with other studies which have used previous uncorrected registered data.

In the field of spatial studies in Iran, only few research is available. These studies have used uncorrected data and are not on national scales in practice. Mahaki et al. studied seven most common cancers in Iran using shared component model and adjusted certain risk factors, identifying northwest, north, and northeast as high risk regions.[25] In the study by Kavousi et al. via space time scan statistics approach, a cluster of high-risk areas were distinguished, namely northwest, north, and some parts in the central provinces.[26] Via ATA poison kriging method, Asmarian et al. (2003-2010) identified the northern and western halves of Iran as high-risk regions.[27]

The maps of the current study after misclassification correction were significantly different from before the correction. The maps revealed that in addition to northern parts, the southern provinces may also be considered as high-risk regions. The eastern provinces, previously classified as high-risk areas (such as r-Khorasan), are currently classified as low-risk areas. Additionally, the western provinces located at low-risk area, were classified as high-risk areas (such as Ilam), also in line with central provinces. Moreover, the risk of GC was clearly underestimated before correction compared with after it. The presence of misclassification problem in the previous uncorrected data may be a reason for such changes.

According to the results of the covariates effect for both genders and after misclassification correction, we found that RR was inversely associated with MHI and directly correlated with UER; however, the association of RR with MHI was significant, while with UER was not

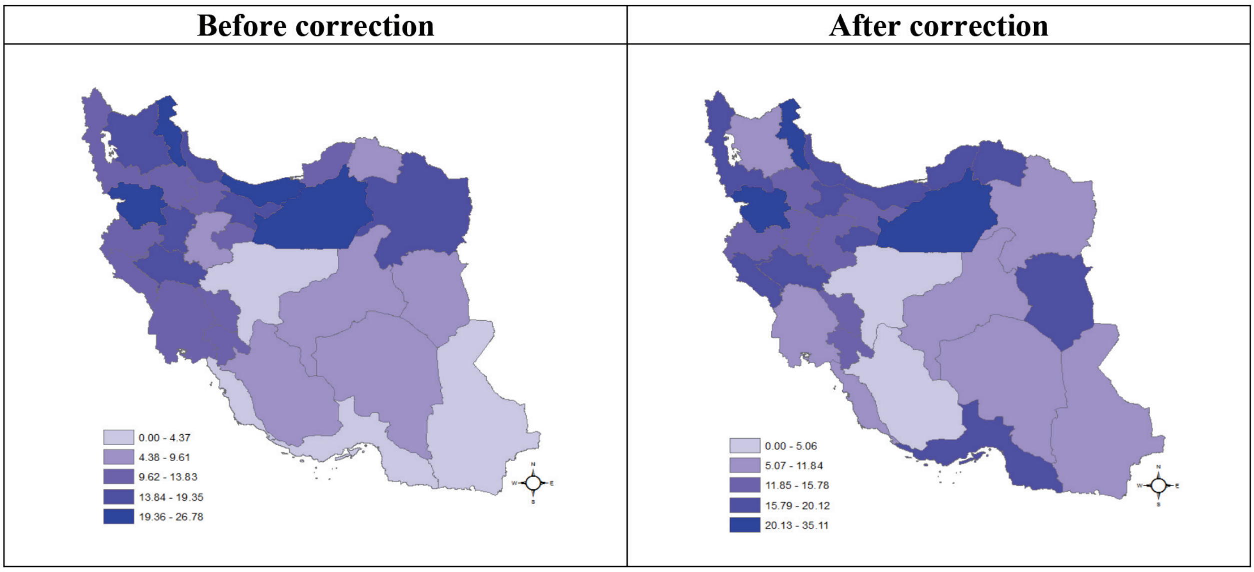

Figure 2. Pattern of age standardized rate for gastric cancer in Iran (for both genders and before or after misclassification correction) in 2008.

Table 3. Summary of posterior estimates for gastric cancer relative risk and socioeconomic risk factors according to misclassification correction status in Iran (2008)

			Mean	SD+	2.5 %	Median	97.5 %
Before misclassification correction	**Women**						
		RR	0.948	0.443	0.115	0.855	1.959
		UER	-0.993*	0.056	-0.204	-0.088	-0.020
		MHI	-0.014*	0.004	-0.021	-0.016	-0.007
	Men						
		RR	0.961	0.470	0.209	0.896	1.923
		UER	0.0007	0.011	0.001	-0.018	-0.002
		MHI	-0.009*	0.001	-0.012	-0.009	-0.006
	Both genders						
		RR	0.968	0.459	0.182	0.866	1.943
		UER	-0.016	0.012	-0.037	-0.012	0.004
		MHI	-0.013*	0.002	-0.018	-0.014	-0.008
After misclassification correction	**Women**						
		RR	1.141	0.500	0.389	1.068	2.431
		UER	-0.056	0.070	-0.188	-0.062	0.094
		MHI	-0.034*	0.017	-0.068	-0.036	-0.006
	Men						
		RR	1.168	0.473	0.339	1.093	2.470
		UER	-0.107	0.063	-0.230	-0.101	0.010
		MHI	-0.004	0.004	-0.013	-0.002	0.001
	Both genders						
		RR	1.129	0.445	0.347	1.076	2.342
		UER	0.006	0.012	-0.012	0.003	0.032
		MHI	-0.008*	0.002	-0.013	-0.009	-0.005

+Standard deviation; *Significant

significant. The result of MHI covariate is consistent with Congdon study on Coronary heart disease in England. In Congdon study, where spatial structural equation modelling was employed, using such covariates led to the introduction of regions with high need for health care.[28]

The strength of our study is that it employed valid corrected data and Bayesian method with regards to the effects of neighboring points, a statistical method increasing the precision of parameter estimation; however, in other procedures, this point usually is neglected. In Iran, the information regarding diet and biologic and genetic risk factors of cancers is not nationally available. Therefore, their effects were not included in the models, which is a limitation on the present research.

Conclusion

The results of the current study indicated how misclassification of registered data can lead to seduction in just health services allocation. Since the northern half and south border of Iran run high risks of GC, preventive measures are seriously considered by health policy makers. Moreover, socioeconomic and environmental GC risk factors and MHI play major roles in the risk of GC. Since a precise introduction of high risk areas is conducive to an accurate allocation of health services, it is important to reduce the misclassification error of the registered data.

Acknowledgement

This study was supported by Shahid Beheshti University of Medical Sciences, Tehran, Iran (grant number 17781).

Conflict of Interest

None declared.

References

1. Forman D, Burley V. Gastric cancer: global pattern of the disease and an overview of environmental risk factors. *Best Pract Res Clin Gastroenterol.* 2006; 20(4):633-49. doi: 10.1016/j.bpg.2006.04.008.
2. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. *CA Cancer J Clin*. 2011;61(2):69-90. doi: 10.3322/caac.20107.
3. Herszenyi L, Tulassay Z. Epidemiology of gastrointestinal and liver tumors. *Eur Rev Med Pharmacol Sci*. 2010;14(4):249-58.
4. Lawson A, Biggeri A, Boehning D. Disease mapping models: an empirical evaluation. Disease Mapping Collaborative Group. *Stat Med.* 2000;19:2217-41.
5. Clarke KC, McLafferty SL, Tempalski BJ. On epidemiology and geographic information systems: a review and discussion of future directions. *Emerg infect Dis.* 1996;2(2):85. doi: 10.3201/eid0202.960202.
6. Cassetti T, La Rosa F, Rossi L, D'Alò D, Stracci F. Cancer incidence in men: a cluster analysis of spatial patterns. *BMC Cancer.* 2008;8(1):344. doi: 10.1186/1471-2407-8-344.
7. Downing A, Forman D, Gilthorpe MS, Edwards KL, Manda SO. Joint disease mapping using six cancers in the Yorkshire region of England. *Int J Health Geogr.* 2008;7(1):41. doi: 10.1186/1476-072X-7-41.
8. Arts DG, Keizer NF, Scheffer GJ. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *J Am Med Inform Assoc.* 2002;6(9):600-11. doi: 10.1197/jamia.m1087.
9. Arts DG, De Keizer NF, Scheffer G-J. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *J Am Med Inform Assoc.* 2002;9(6):600-11.
10. Burnet N, Jefferies S, Benson R, Hunt D, Treasure F. Years of life lost (YLL) from cancer is an important measure of population burden—and should be considered when allocating research funds. *Br J Cancer.* 2005;92(2):241. doi: 10.1038/sj.bjc.6602321.
11. Hajizadeh N, Pourhoseingholi MA, Baghestani AR, Abadi A, Ghoreshi B. Years of life lost due to gastric cancer is increased after Bayesian correcting for misclassification in Iranian population. *Gastroenterol Hepatol Bed Bench.* 2016;9(4):295.
12. Edwards JK, Cole SR, Chu H, Olshan AF, Richardson DB. Accounting for outcome misclassification in estimates of the effect of occupational asbestos exposure on lung cancer death. *Am J Epidemiol.* 2013;179(5):641-7. doi: 10.1093/aje/kwt309.
13. Taghavi A, Fazeli Z, Vahedi M, Baghestani AR, Zali MR, Pourhoseingholi MA. Pancreatic cancer mortality and misclassification--bayesian analysis. *Asian Pac J Cancer Prev.* 2011;12(9):2271-4.
14. Aghajani H, Eatemad K, Goya M, Ramezani R, Modirian MNF, Nadali F. Iranian annual of national cancer registration report 2008-2009. *Center for Disease Control.* 2011.
15. Hajizadeh N, Pourhoseingholi MA, Baghestani AR, Abadi A, Zali MR. Bayesian adjustment for over-estimation and under-estimation of gastric cancer incidence across Iranian provinces. *World J Gastrointest Oncol.* 2017;9(2):87.
16. Wakefield J. Disease mapping and spatial regression with count data. *Biostatistics.* 2006;8(2):158-83. doi: 10.1093/biostatistics/kxl008.
17. Bernardinelli L, Montomoli C. Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. *Stat Med.* 1992;11:983-1007.
18. Besag J, York J, Mollie A. Bayesian image restoration with two applications in spatial statistics. *Ann Inst Stat Math.* 1991; 43: 1–59. doi: 10.1007. BF00116466.
19. Kavousi A, Meshkani MR, Mohammadzadeh M. Spatial analysis of relative risk of lip cancer in Iran: a Bayesian approach. *Environmetrics*. 2009;20(4):347-59.
20. Zayeri F, Kavousi A, Najafimehr H. Spatial analysis of Relative Risks for skin cancer morbidity and mortality in Iran, 2008-2010. *Asian Pac J Cancer Prev.* 2015;16(13):5225-31.
21. Rytkönen MJ. Not all maps are equal: GIS and spatial analysis in epidemiology. *Int J Circumpolar Health.* 2004;63(1):9-24. doi: 10.3402/ijch.v63i1.17642.
22. Kaboli SA, Pourhoseingholi MA, Moghimi-Dehkordi B, Safaee A, Habibi M, Pourhoseingholi A, et al. Factors associated with functional constipation in Iranian adults: a population-based study. *Gastroenterol Hepatol Bed Bench.* 2010;3(2): 83-90.
23. Pourhoseingholi MA, Faghihzadeh S, Hajizadeh E, Abadi A, Zali MR. Bayesian estimation of colorectal cancer mortality in the presence of misclassification in Iran. *Asian Pac J Cancer Prev.* 2009;10(4):691-4.
24. Hajizadeh N, Pourhoseingholi MA, Baghestani A, Abadi A, Zali MR. Under-estimation and over-estimation in gastric cancer incidence registry in Khorasan Provinces, Iran. *The Cancer Press.* 2016;2(3):43-5.
25. Mahaki B, Mehrabi Y, Kavousi A, Akbari ME, Waldhoer T, Schmid V, et al. Multivariate disease mapping of seven prevalent cancers in Iran using a shared component model. *Asian Pac J Cancer Prev.* 2011;12:2353-8.
26. Kavousi A, Bashiri Y, Mehrabi Y, Etemad K, Teymourpour A. Identifying high-risk clusters of gastric cancer incidence in Iran, 2004-2009. *Asian Pac J Cancer Prev.* 2014;15(15):10335-7.
27. Asmarian N, Jafari-Koshki T, Soleimani A, Ayatollahi SMT. Area-to-area poisson kriging and spatial bayesian analysis in mapping of gastric cancer Incidence in Iran. *Asian Pac J Cancer Prev.* 2016;17(10):4587-90.
28. Congdon P. A spatial structural equation model with an application to area health needs. *J STAT Comput Simul*. 2010;80(4):401-12.