82283_Auto_Edited.docx

WORD COUNT 5665
TIME SUBMITTED 12-APR-2023 05:07PM
PAPER ID 98636036

Name of Journal: *Artificial Intelligence in Gastroenterology*

Manuscript NO: 82283

Manuscript Type: ORIGINAL ARTICLE

Retrospective Study

Risk factor profiles for gastric cancer prediction with respect to Helicobacter pylori: A study of a tertiary care hospital in Pakistan

Shahid Aziz, Simone König, Muhammad Umer, Tayyab Saeed Akhter, Shafqat Iqbal, Maryum Ibrar, Tofeeq Ur-Rehman, Tanvir Ahmad, Alfizah Hanafiah, Rabaab Zahra, Faisal Rasheed

Abstract

BACKGROUND

Gastric cancer (GC) is the fourth leading cause of cancer-related deaths worldwide. Diagnosis relies on histopathology and the number of endoscopies is increasing. *Helicobacter pylori* infection is a major risk factor.

AIM

The present study was aimed at developing an *in-silico* GC prediction model to reduce the number of diagnostic surgical procedures. The meta-data of patients with gastroduodenal symptoms, risk factors associated with GC, and *H. pylori* infection status from Holy Family Hospital Rawalpindi, Pakistan, were used with machine learning.

METHODS

A cohort of 341 patients was divided into three groups (normal gastric mucosa - NGM, gastroduodenal disorders - GDD, and GC). Information associated with socioeconomic and demographic conditions and GC risk factors was collected using a questionnaire. *H.*

pylori infection status was determined based on urea breath test. The association of these factors and histopathological grades was assessed statistically. K-Nearest Neighbors (KNN) and Random Forest (RF) machine learning models were tested.

RESULTS

This study reported an overall frequency of 64.2% (219/341) of *H. pylori* infection among enrolled subjects. It was higher in GC (74.2%, 23/31) as compared to NGM and GDD and higher in males (54.3%, 119/219) as compared to females. More abdominal pain (72.4%, 247/341) was observed than other clinical symptoms including vomiting, bloating, acid reflux and heartburn. The majority of the GC patients experienced symptoms of vomiting (91%, 20/22) with abdominal pain (100%, 22/22). The multinomial logistic regression model was statistically significant and correctly classified 80% of the GDD/GC cases. Age, income level, vomiting, bloating and medication had significant association with GDD and GC. A dynamic RF GC-predictive model was developed, which achieved >80% test accuracy.

CONCLUSION

GC risk factors were incorporated into a computer model to predict the likelihood of developing GC with high sensitivity and specificity. The model is dynamic and will be further improved and validated by including new data in future research studies. Its use may reduce unnecessary endoscopic procedures. It is freely available.

INTRODUCTION

Gastric cancer (GC) is the fourth most common cancer in the world and the second-most common cause of cancer-related death [1] with the highest incidence observed in Eastern Asia and the lowest in Western Europe and North America [2]. The main environmental factor causing GC is *Helicobacter pylori* infection [1], and it has been classified as a class I carcinogen by the International Agency for Research on Cancer [3]. It is, however, an insufficient cause, and other hereditary [4], environmental and lifestyle

factors are of importance in GC development as well [1, 5-8]. GC risk factors and epidemiology in Pakistan were reviewed in 2015 [9] and 2018 [10] stressing the importance of sanitary conditions, purified drinking water and healthy nutrition in a developing country with 24.3% poverty rate [11]. The latter meta-analysis remarked on the population heterogeneity in different parts of the country, where various ethnic groups follow their own lifestyle traditions so that cancer statistics vary considerably [10]. A National Cancer Registry is presently not available but is in the process of being set up by the Pakistan Health Research Council [11].

GC risk factors include age [12], gender [13] and all factors which are commonly named as general health risks such as smoking [14,15], alcohol and junk food consumption as well as reduced physical exercise [5,6,16,17]. Diet and, in particular, controlled sugar and salt consumption play a specific role in GC prevention [18-20]. Proton pump inhibitors (PPI), which are routinely prescribed in the management of gastric-acid-related disorders, may also pose a risk, when improperly used [21, 22]. Harvard University adds in its "10 commandments of cancer prevention" [23] factors such as exposure to radiation and industrial and environmental toxins, little sleep and lack of vitamin D to the list. Furthermore, local habits in different countries or ethnicities may influence the risk of GC development. In Asia, for instance, Miswak (toothbrush tree, *Salvadora persica* L.) is commonly used for oral hygiene counteracting *H. pylori* infection [24]. High chili consumption in some regions of South America, on the other hand, sensitizes the mucosa and poses a cancer risk [25].

Histological examination of gastric biopsies is currently the gold standard for GC diagnosis [16]. However, the demand for endoscopy is increasing along with the financial burden for the health care system so that the number and appropriateness of referrals is more and more discussed [26]. Guidelines were published in what instances endoscopic biopsies should be performed [27], not only for economic reasons, but also to avoid stressing patients with false-positive results in cases of abnormal appearance of gastric mucosa in endoscopy but normal histopathology [28]. Moreover, health care-allied infections are significantly associated with contaminated endoscopes. The most

commonly used flexible multi-channel endoscopes need utmost care in high-level disinfection and proper cleaning before endoscopic procedures, as they cannot be heat-sterilized. Otherwise, bacteria may form biofilms on the inner surfaces and pose a serious risk to patients [29].

In the Center for Liver and Digestive Diseases of the tertiary care Holy Family Hospital in Rawalpindi we have also seen an overload in referrals to endoscopic procedures. In order to find a measure for improved patient referral we collected clinical data of 341 patients having symptoms of gastroduodenal disorders and asked them to fill in a questionnaire concerning their living conditions as well as diet and daily habits. It included the risk factors discussed above and factors important with respect to *H. pylori* infection like overcrowding and source of drinking water, because sanitary conditions contribute significantly to the spread of microorganisms [30-32]. The aim of this study was to set up an *in silico*-model, which could be continuously trained with new patients of our clinic, and which would allow us to limit the referrals to endoscopy to the most serious cases based on risk factor assessment. Such machine-learning models are increasingly being used in gastroenterology [33-35], most recently for the prediction of GC risk after *H. pylori* eradication [35]. All these efforts were, however, retrospective studies, while we try to build up a prognostic tool, which is closely associated with the clinic and integrated in everyday use, and which is constantly being improved with new data. Despite the low number of starting data - in comparison to these other models, which are in part based on ten thousands of patients, we can present a model, which already predicts the GC risk with an impressive >80% confidence.

Artificial intelligence (AI) is playing an increasing role in the healthcare industry including gastroenterology and gastrointestinal oncology. AI can assist physicians in invasive procedures such as endoscopy, capsule endoscopy, and colonoscopy for disease diagnosing [64], radiology [65], and the detection of the cancerous and precancerous lesions in the intestine [66].

MATERIALS AND METHODS

2.1 Ethical approval and study population

Ethical approvals were granted from the Ethical Technical Committee, Pakistan Institute of Nuclear Science and Technology (PINSTECH), Islamabad (Ref.-No. PINST/DC-26/2017), the Bioethics Committee, Quaid-i-Azam University, Islamabad, Pakistan (Ref.-No. BBC-FBS-QAU2019-159), and the Institutional Research Forum, Holy Family Hospital, Rawalpindi Medical University, Rawalpindi (Ref.-No. R-40/RMU).

Inclusion and exclusion criteria

Primary data of 341 patients having persistant dyspeptic symptoms of gastroduodenal disorders including acid reflux, abdominal pain, heartburn, vomiting, and bloating, or alarm symptoms who were thus attending the Centre for Liver and Digestive Diseases, Holy Family Hospital, Rawalpindi for upper gastroduodenal endoscopy of age group above 18 years was collected in this study from 2018 to 2021. They also signed the informed written consent.

However, patients having a history of confounders of gastric cancer such as gastric surgery, corrosive intake, varicel bleed with chronic liver disease, or use of antibacterial and gastric acid inhibitors during the past 30 days which may affects on diagnosis of *H. pylori* infection and anticancer drugs were excluded from this study, so were pregnant women.

After diagnostic endoscopic evaluation, the enrolled patients were divided into three groups: normal gastric mucosa (NGM), GC and gastroduodenal diseases (GDD). The GDD group included patients who had gastritis (mild, moderate, marked and PAN gastritis (chronic form of gastritis, which affects the entire gastric mucosa). The patients with gastritis were subcategorized into mild (mild erythma or scanty erosions), moderate (neither mild nor marked), and marked (diffuse erythma, nodularity, hypertrophy of gastric folds and friability of gastric mucosa) according to Kyoto classification system [67]. Moreover, ulcers (gastric, duodenal, and peptic ulcer diseases) were also included in this group.

Questionnaire for exploring demographics and socioeconomic status

Patients were interviewed using a Likert-scale questionnaire developed earlier for the investigation of *H. pylori* infection in Pakistan [30]. Information associated with socioeconomic and demographic conditions such as gender, age, education, income, and living conditions was collected in addition to GC risk factors including specific dietary habits. There have been studies, which associated dairy products with GC [36] and those who did not [37] as well as studies, which evaluated the influence of red and processed meat [38], high salt consumption due to salted fish and meat [20], and black and green tea [7]. An unhealthy diet very high in carbohydrates (rice, potato) and low in fresh vegetables and fruit is also critical [1,4,7,8] and questions to that effect were included in the questionnaire. Moreover, the history concerning the intake of antibacterial drugs, proton pump inhibitors (PPI), non-steroidal anti-inflammatory drugs (NSAIDS) and other medicines was recorded. Categories of responses were defined as listed in Table 1.

Diagnosis of *H. pylori* infection

Standard non-invasive and invasive diagnostic tests were performed for the determination of *H. pylori* infection. All the modalities including nuclear stable isotope ^{13}C urea breath test (UBT), histopathological examinations (HPE) and rapid urease test (RUT) were used to diagnose *H. pylori* infection with the exception that biopsy specimens were not available for all the patients. The ^{13}C UBT was, however, used for all enrolled subjects.

2.4.1 Nuclear Stable Isotope ^{13}C Urea Breath Test (UBT)

Active *H. pylori* infection was determined using non-invasive nuclear stable isotope ^{13}C UBT as described previously [30]. Briefly, after all-night fasting, a pre-dose breath sample was collected from the patient. A dose containing 75 mg ^{13}C enriched urea (Cambridge Isotope Laboratories, USA) was given to the patient and post-dose breath sampling was performed after 30 min. Breath samples were analyzed for $^{13}CO_2/^{12}CO_2$ ratio using BreathMAT^{plus} mass spectromter (Thermo Finnigan, Germany) and Delta V Plus mass spectrometer (Thermo Scientific, USA). A change in the δ ^{13}C value over baseline of more than 3‰ was considered positive.

2.4.2 Gastric biopsy collection

Specimens were collected from those patients who had symptoms suggestive of a need for upper gastroduodenal endoscopy. Multiple biopsy specimens were collected from antrum and corpus within 3 cm of the pylorus of each patient undergoing this surgery. Biopsy specimens were placed in 10% formalin for HPE. One biopsy was collected for RUT.

2.4.3 RUT

The rapid urease kit to assess the active growth of *H. pylori* was indigenously prepared in Patients Diagnostic Lab, PINSTECH. Briefly, fresh gastric biopsy specimen were immediately placed in urea agar base with 40% urea solution for 1 h of incubation at 37 °C. A change of color from pale yellow to pink red was interpreted as a positive result.

2.4.4 HPE

Gastric (antrum and corpus) biopsy specimens were processed for histopathological examination according to the Operative Link for Gastritis Assessment (OLGA/OLGIM) scoring [40] alongside with Lauren and WHO classification systems [41] for the determination of NGM, gastritis, gastric ulcer, duodenal ulcer and GC differentiation and invasions.

2.5 Statistical analysis

Chi-squared (χ^2) test was used to assess the association of socioeconomic demographics, different risk factors, and histopathological grades among the three groups (NGM, GDD, GC). Spearman correlation coefficient test was employed to find the relationship between *H. pylori* infection and histopathological variables among gastric biopsies of antrum and corpus. The association between the predictor variables in the three groups was evaluated using multinomial logistic regression analysis. Nine variables having a p value <0.1 were selected for multinomial logistic regression analysis. Risk factors included in the multivariable model were age, education level, income level, symptoms (abdominal pain, acid reflux, vomiting, bloating), chili consumption, excessive intake of salt and medication usage. Frequency categories were

combined to achieve sufficient statistical power. Multinomial logistic regression analysis was used to determine factors associated with the three groups. To evaluate the interaction of different risk factors among the three groups, likelihood ratio tests were used to calculate *p* values comparing models with main effects to models with main effects plus relevant interaction terms. PCA was carried out for risk factors, symptoms and *H. pylori* tests restricting the number of factors to three. For initial data classification with respect to endoscopic data and a focus on GC, decision tree analysis was performed with risk factors. All *p* values were reported as two-sided test with an alpha level of 0.05. Statistical analysis was carried out with SPSS 21.0 statistical software (SPSS Inc, Chicago, USA).

RESULTS

3.1. General characteristics of study participants

Participants (341) with the mean age of 41.9 ±15.9 years and an age range from 18 to 87 years were included in this study. All data are supplied in Supplementary Table 1. The overall frequency of *H. pylori* infection was 64.2% (219/341). The enrolled patients were separated in the following groups: NGM 15% (50/341), GC 9.1% (31/341), and GDD 76.2% (260/341). The frequency of *H. pylori* infection among NGM participants was 72% (36/50), 62% (160/260) in GDD, and 74.2% (23/31) in GC. About half of the participants were male (177/341, 51.9%); 48.1% (164/341) were females. The frequency of *H. pylori* infection was higher in males (54.3%, 119/219) as compared to females (45.7%, 100/219). Clinical symptoms observed among enrolled patients were abdominal pain (72.4%, 247/341), vomiting (57.8%, 197/341), bloating (54.5%, 186/341), acid reflux (52.8%, 180/341) and heartburn (52.8%, 180/341). The majority of the GC patients were older than 45 years (71%, 22/31) and experienced symptoms of vomiting (91%, 20/22) with abdominal pain (100%, 22/22).

Descriptive characteristics of the cohort and results of the Chi-squared (χ^2) test to assess the association of socioeconomic demographics, risk factors, and histopathological grades among the three groups (NGM, GDD, GC) are presented in

Table 2. Significant factors were age, education (one-third of the participants were illiterate) and, conclusively, income level, and the clinical symptoms (except heartburn). Cross-correlation was computed for visualization of the data set as is exemplary shown for age, gender and rapid urease test (RUT) results in Figure 1 and Supplementary Table S1.

Multinomial logistic regression analysis

The associations of risk factors with GDD and GC among the three groups are presented in Table 3. Chi squared analysis showed a significant association at $p < 0.05$ between 7 independent variables among 3 groups. Out of 38 indicators, 9 variables added to the multinomial logistic regression analysis with $p < 0.1$. Multinomial logistic regression was performed to ascertain the effects of predictor variables on the likelihood that participants had GDD or GC. Model fitting information described the relationship between the dependent and independent variables and revealed that the probability of the model Chi-square 97.028 was 0.01, less than the level of significance of 0.05 (i.e., $p < 0.05$). The model explained 32.0% (Nagelkerke $R2$) of the variance in groups and correctly classified 80% of the cases; 10% of the cases from GC, 98% from GDD and 30% of the NGM participants.

According to Wald statistics, age, income level, vomiting, bloating and medication were the significant factors associated with GDD and GC. People younger than 45 years were less likely to have GC as compared to GDD (OR 0.19, 95%CI 0.08-0.46, $p < 0.05$) and as compared to normal (OR 0.08, 95%CI 0.02-0.29, $p < 0.05$). People belonging to the middle class were more likely to have GDD (OR 2.32, 95%CI 1.09-4.91, $p < 0.05$) and GC (OR 4.86, 95%CI 1.25-18.84, $p < 0.05$) as compared to NGM. Similarly, patients without the symptoms of vomiting (OR 0.16, 95%CI 0.05-0.53, $p < 0.05$) and abdominal pain (OR 0.17, 95%CI 0.04-0.72, $p < 0.05$) were less likely to have GC than NGM. Patients without the symptoms of bloating are also less likely to have GDD as compared to NGM (OR 0.37, 95%CI 0.17-0.8, $p < 0.05$) and GC as compared to GDD (OR 0.29, 95%CI 0.1-0.8, $p < 0.05$).

3.3 Upper gastroduodenal endoscopic evaluation

The total of 341 patients underwent upper gastroduodenal endoscopy. Among these patients, 15% (50/341) had NGM, 67% (230/341) patients had gastritis, 9% (30/341) had gastroduodenal ulcers including gastric ulcers (70.0%, 21/30), duodenal ulcers (20%, 6/30), and peptic ulcer disease (10%, 3/30). Those patients with gastric ulcers, duodenal ulcers and peptic ulcer disease had a frequency of *H. pylori* infection 62% (13/21), 83% (5/6) and 67% (2/3), respectively. Moreover, all ulcers were categorized as clean-based ulcers and classified as Forrest III (lesions without active bleeding). Additionally, 9% (31/341) patients were suspected (based on lesion, polyp, and large growth) for GC and their gastric biopsy specimens were taken for histopathological examination (HPE) to rule out the malignancies.

GC evaluation and differentiation

HPEs showed that 51% (117/230) of the patients had mild gastritis, 40% (93/230) moderate gastritis, and 2% (4/230) marked gastritis. The frequency of *H. pylori* infection in patients with mild gastritis was 62% (72/117), with moderate gastritis 59% (55/93), and with marked gastritis 0.5% (2/4). A total of 31 patients were histopathologically confirmed for GC. Among those patients, 23% (7/31) had first and 77% (24/31) had advanced stage GC. The frequency of *H. pylori* infection in first and advanced stage GC was 86% (6/7) and 71% (17/24), respectively. Additionally, those patients were also evaluated and differentiated into various cancer types including adenocarcinoma (48%, 15/31), signet ring cell carcinoma (45%, 14/31) and undifferentiated carcinomas (6.4%, 2/31) with 93% (13/14), 60% (9/15) and 50% (1/2) frequency of *H. pylori* infection, respectively. Moreover, gastric biopsies were also examined and graded according to Lauren and WHO classifications into intestinal (19%, 6/31), diffuse (81%, 21/31), tubular (48%, 15/31) and poorly cohesive (52%, 16/31) carcinomas. The frequency of *H. pylori* infection among these patients was: 33% (2/6), 68% (21/31), 60% (9/15), 88% (14/16), respectively.

3.4.1 Correlation of histopathological variables of antrum and corpus biopsies

The Spearman coefficient correlation test for histopathological assessment of multiple gastric biopsies from antrum and corpus revealed a highly significant correlation ($p < 0.05$) between *H. pylori* infection and histopathological grades including *H. pylori* load, neutrophil infiltration, mononuclear cell infiltration, inflammation, atrophy, atypia, metaplasia, dysplasia, atrophy score (OLGA), metaplasia score (OLGIM), gastritis and ulceration (Table 4).

3.5 Principal components analysis (PCA) and decision trees

When testing for the factors with the most influence in the dataset using PCA, not unexpectedly, factors related to *H. pylori* infection (^{13}C UBT, RUT) were dominant followed by characteristic symptoms for gastroduodenal diseases (heartburn, vomiting, reflux; Supplementary Table S2). Decision tree analysis with a focus on GC (Supplementary Figure S1) revealed age as the main separator with people younger than 50 years showing only 1/3 of all GC cases. When age was excluded from the analysis (Figure S1B), the factor abdominal pain collected 28 of 31 GC patients in the node, which were further split for 26 suffering from vomiting. Bloating was not a useful selection criterion for GC, because only 1/3 of all GC cases reported it.

3.6 Machine-learning algorithm

Resulting from extensive literature review and the findings of this study, 23 factors associated with GC were selected and used to train a GC prediction model using python language (Table 5). The diagnostic approach using machine learning was carried out in two steps, firstly model trained itself by recognizing patterns in the data of all classes of gastric diseases and secodly, the pre-learned model classified new patients after identification of similar pattern of newly provided data. The probabilities of specific disease were predicted due to closer pattern after input of patients data.

The primary dataset (parameters in textual and structural format, Supplementary file Training_Testing_Data) contained upper-gastroduodenal symptoms, potential GC risk factors, *H. pylori* infection status, and clinical endoscopic and histopathological

findings. Factor categories were reduced to yes and no in some cases to provide sufficient numbers of samples, respectively, analysis power. The primary data was imbalanced containing a higher number of gastritis patients as compared to ulcer and GC patients. Therefore, 70% samples of each class were used to train the model and the remaining 30% for testing (Table 6). The algorithm randomly performed this 70-30 distribution of the dataset. During testing, the pre-learned machine learning model truly classified 72% cases of each class with greater accuracy.

Two machine learning models based on K-Nearest Neighbors (KNN) and Random Forest (RF) supervised learning algorithms were separately trained to calculate the risk of a specific gastroduodenal disease. In the KNN model, a simple elucidation distance of the test samples with all training samples was calculated. Top 'K' training samples, i.e. patient feature vectors with a minimum distance with the test samples, decided the highest risk of a certain disease by voting for the most frequent class. In general, for samples in n-dimensional Euclidean space, the distance is, with p and q being two points in Euclidean n-space:

RF is an ensemble of, in our case 10, decision trees. It eradicated the over-fitting that is a major issue of decision tree. Each tree made decisions based on importance of each risk factor, i.e., starting from features that are more distinct to the less important features. Importance is defined as the distinguishability of a feature and it was measured by Gini Gain or Importance Gain (for more details see Explanation S1 in the Supplementary Data). We have used the Gini Index to train our model. With KNN we achieved 74% and with RF 82% test accuracy. We thus incorporated the latter algorithm in the published software tool. RF is a decision tree based stacking classifier which is freely available with a few tunable hyper parameters. It is not constructed from scratch but trained by using patient's data and also optimized by fine tuning of the important parameters.

The user interface of the GC Prediction System (GCPS) is shown in Figure 2. The input is limited to the most critical factors with respect to risk modelling. The software was written for Windows 10 and is distributed as archive containing an executable program file (link: www.medizin.uni-muenster.de/cu-proteomics/projekte.html). Running the tool simply requires to unzip and join the three archives and then run the executable file on any Windows-based computer. Results are reported online and are saved in pdf-format in the program directory. Via the input page, data can be added to the model to train it further, but this needs to be done in the original python-based environment and is thus not available to the standard user. The source code is shared in collaborations.

Figure 2. Exemplary input to GC prediction tool interface to record patient data, symptoms, ^{13}C UBT results and risk factors. Following input, a click on the "Result" button shows the probability of developing GC. A report can be generated in pdf-format. The "Update Data" button is used only when including new patient data into the model.

DISCUSSION

H. pylori infection is a serious public health problem with a high frequency among the population of developing countries [45]. Globally, 4.4 billion individuals have been identified to harbor *H. pylori*. The frequency of *H. pylori* infection in developing and developed countries has been reported as 70-90% and 10-30%, respectively [46]. Our previous study showed more than 70% frequency of *H. pylori* infection in the northern region of Pakistan [47]. Six years later [48], active *H. pylori* infection was detected in 50% of the symptomatic patients in Pakistan of whom 76% had clinical symptoms like abdominal pain. In the present investigation, we found 64% infection in symptomatic patients indicating a considerable increase over time. As the consistent presence of *H. pylori* infection in a large part of the population provides the basis for several gastroduodenal clinicopathological conditions including gastritis, ulcers and most

importantly GC [1, 3], this is an alarming situation. In earlier studies conducted on symptomatic patients from Pakistan, GC frequency was reported as 6.0% and 6.4%, respectively [10, 49], while, here, 9.1% were calculated. In agreement with our previous findings [48], the infection rate in males (54%) was marginally higher compared to females (46%) possibly due to their higher social interaction in Pakistan. Likely for the same reason, people younger than 46 years were more often infected by *H. pylori* (64%). Infection takes place in childhood and adolescence and reaches its peak in adulthood at an age of 35-44 years [46, 50].

The increased risk of *H. pylori* positivity in developing countries has been associated with several environmental factors including lower socioeconomic conditions such as crowded households and poor hygiene [51]. Already in our previous study [30], these risk factors, further including pets and other household animals, have been significantly associated with *H. pylori* infection. Here, we also showed the influence of education and income level. Educated people can take advantage of the available knowledgebase and better care for their health. Moreover, with education comes job advancement and improved financial means to provide for optimal living conditions. The frequency of *H. pylori* infection (64%) was expectedly higher in patients with comparatively low family income (11,000-30,000 PKR; 51-138 USD, 1 USD = PKR) where living conditions are difficult. About 256,465 PKR (1194 USD) are required for appropriate living conditions and fulfillment of basic needs [52].

Personal hygiene of the oral cavity is another risk factor as the mouth is the first pool of *H. pylori* infection and has a positive correlation with gastroduodenal pathologies [53]. Miswak has been traditionally used in Pakistan for oral hygiene due its antibacterial properties against both Gram positive and negative bacteria [24]. As is demonstrated in this study, a higher risk of *H. pylori* infection was found in patients who did not use it or other forms of oral hygiene.

Dietary habits such as meat consumption and the use of outdoor potable water were described as significant independent variables for both *H. pylori* infection and GC risk before [54]. A study conducted in Korea indicated that high salt intake was

associated with a higher risk of atrophic gastritis and intestinal metaplasia [55] and other authors showed that it can lead to the onset of pre-malignant lesions [56]. In addition, the carcinogenic effects of major *H. pylori* virulence factor cytotoxin associated gene A (*cagA*)-positive strains were increased [57, 58]. We confirmed the higher risk of *H. pylori* infection (73%) in patients with a higher salt intake than 5 g/day as recommended by the WHO [44].

A diet rich in carbohydrates and sweets is generally not healthy and the positive correlation with *H. pylori* infection was established in a study conducted in Japan in 2016 [59] as well as here. It was also reported that for people who engage in regular exercise in the presence of *H. pylori* infection, the GC risk was reduced by approximately 50% in both males and females [60]. We saw more *H. pylori* infections in patients who did not have a habit of physical activity in their routine life but there was no correlation with GC incidence.

It has been suggested that a *Lactobacillus rhamnosus*-providing dairy-rich diet may counteract *H. pylori* infection [36]. In general, dairy products are a source of many nutrients and are highly recommended in dietary guidelines. Nevertheless, some studies found adverse effects of dairy consumption with GC [37] that is why we included this factor in our questionnaire. No clear conclusion can however be drawn from the available reports as some studies appear to have been flawed in their design [37]. Given the clear advantages of diet containing milk and dairy products, we do not wish to over interpret our data, which positively correlate *H. pylori* infection and use of dairy products. It may rather be advisable for patients sensitive to gastroduodenal symptoms to test their response to milk and other dairy products (allergies) and adjust their diet accordingly.

Black and green tea have been named as GC risk factors [7], because, in particular, green tea contains antioxidant compounds, which showed remarkable antibacterial activity especially against *H. pylori and were beneficial against associated gastric diseases during in vitro* and *in vivo experiments [61]. As did other authors [62], we observed more H. pylori* infection in patients who did not drink green tea in their routine life *(68%).*

Clinical symptoms such as vomiting were significant independent variables, which matched the results of others [51]. The coefficient correlation for *H. pylori* loads (0.542), neutrophil (0.644) and mononuclear cell infiltration (0.173) for antrum and corpus was assessed with a significance level of $p = 0.000$ before [50]. In our study, there was also a significant positive correlation ($p < 0.01$) among histopathological grades including *H. pylori* load (0.991), neutrophil (1.000) and mononuclear cell infiltration (0.942) for antrum and corpus biopsies. The significant correlation among all histopathological grades in gastric biopsies suggests that a minimum number of biopsies can be sufficient to rule out malignancies. Other authors have reported the need for 6-8 gastric biopsies to ensure confident diagnosis [63]. A high number of gastric biopsy specimens may, however, create problems apart from procedure prolongation including active bleeding [63].

We have incorporated the pre-endoscopic patient's data from this study and the literature for risk factors and *H. pylori* infection status into a machine-learning algorithm and generated a GC model, which the practitioner can use for a quick check of the GC risk. Other efforts with respect to computer models in gastroenterology were retrospective studies [33-35], while we aim for a prognostic tool, which is constantly being improved with new data. Our model reached >80% confidence in GC prediction and it may be helpful in making a decision pro and con gastroduodenal endoscopy in some cases. However, it is only based on 341 patients of which 31 had GC, so it clearly cannot be used as sole decisive factor; the experience of the physician is not to be underestimated. We plan to continuously improve the tool by the addition of new patient data from our clinic. We will release an updated version to the scientific community from time to time, because we do believe that this screening tool can be helpful.

CONCLUSION

We report a high and increasing level of *H. pylori* infection in Pakistan and its association with different risk factors, which, in turn, have direct or indirect

relationships with gastroduodenal diseases including gastritis, ulcers, and GC. Our study identified GC risk factors such as age, sanitary conditions and clinical symptoms and incorporated them into a dynamic computer tool for GC prediction.

GC is a huge burden in developing countries. Awareness should be raised at an individual level through social media, schools, medical camps, and other means of public education to reduce the risk of gastric malignancies especially in the presence of *H. pylori* infection. Individual habits regarding diet or hygiene can be targeted in that way. Other risk factors require political intervention or governmental decisions. *H. pylori* infection monitoring and eradication strategies, for instance, are means of GC prevention [54]. The general improvement of living conditions and infrastructure will advance sanitary conditions and, conclusively, support the battle against GC. The investigation assists the healthcare authorities in their understanding of the burden of GDD and GC, which is intertwined with *H. pylori* infection.

ARTICLE HIGHLIGHTS

Research background

Gastric cancer is the 4th main reason for cancer-associated deaths around the globe. Diagnosis mainly depends on histopathological examinations and the number of endoscopic procedures is increasing. *Helicobacter pylori* infection is a main risk factor for this cancer.

Research motivation

The increasing prevalence of gastric cancer due to late diagnosis or at an advanced stage was the main cause to conduct this research study to diagnose gastric cancer at an early stage.

Research objectives

The main research objectives of this study were;

1. Diagnosis of H. pylori infection

2. Development of gastric cancer prediction model using non-invasive characteristics of enrolled subjects

Research methods

The 341 dyspeptic patients were enrolled after endoscopic evaluation and metadata was collected using a Likert scale questionnaire. The infection status was determined with the help of three modalities including 13C UBT, rapid urease test, and histopathological examinations. The RF-gastric cancer prediction model and developed using non-invasive characteristics of patients.

Research results

This study reported a higher frequency of *H. pylori* infections among enrolled subjects. It was greater in gastric cancer as compared to other groups and also higher in males in comparison with females. The abdominal pain was observed more than other clinical symptoms. The majority of gastric cancer patients experienced symptoms of vomiting with abdominal pain. The multinomial logistic regression model correctly classified 80% of gastric cancer cases. The RF GC predictive model achieved >80% test accuracy.

Research conclusions

The gastric cancer risk factors were incorporated into a computer model to predict the likelihood of developing gastric cancer with high sensitivity and specificity. The model is dynamic and will be further improved and validated by including new data in future research studies. Its use may reduce unnecessary endoscopic procedures.

Research perspectives

The computer model will predict the likelihood of developing gastric cancer with high sensitivity and specificity. Moreover, It will also be helpful in diagnosing other

gastric diseases such as gastritis and ulcer and at least assist gastroenterologists to start palliative therapy to reduce unnecessary endoscopic procedures.

ACKNOWLEDGEMENTS

We acknowledge the contributions and expertise of Dr Aiza Saadia (Department of Histopathology, Army Medical College, Rawalpindi, Pakistan.) during histopathological examinations of gastric biopsies. Additionally, we appreciate the technical assistance of Tariq Mehmood and Kashif Siddique (Patients Diagnostic Lab, PINSTECH, Isalamabad, Pakistan) during this research project. Moreover, we also are grateful for the efforts of technical staff (Irfana Danish, Kiran Shamim, Sajjad Ahmad, Farhat Mehmood, and Muhammad Majid) from endoscopic center of Holy Family Hospital, Rawalpindi, Pakistan in collecting gastric biopsies during endoscopic procedures.

82283_Auto_Edited.docx

ORIGINALITY REPORT

6%
SIMILARITY INDEX

PRIMARY SOURCES

1	www.ncbi.nlm.nih.gov Internet	129 words — 2%
2	prr.hec.gov.pk Internet	44 words — 1%
3	link.springer.com Internet	39 words — 1%
4	Maxim Norkin, Lakshmikanth Katragadda, Fei Zou, Sican Xiong et al. "Minimal residual disease by either flow cytometry or cytogenetics prior to an allogeneic hematopoietic stem cell transplant is associated with poor outcome in acute myeloid leukemia", Blood Cancer Journal, 2017 Crossref	34 words — 1%
5	Qurat ul an Sabir, Uzma Yasmeen, Muhammad Hanif, Tri Nguyen-Quang. "IDENTIFICATION OF INFLUENCE OF THE RISK FACTORS TO ALGAL BLOOM IN FRESH WATER RESERVOIRS USING MULTINOMIAL LOGISTICS REGRESSION", Advances and Applications in Statistics, 2020 Crossref	33 words — 1%
6	www.researchgate.net Internet	17 words — < 1%

7	Benedikt Weber, Elias Marquart, Julia Deinsberger, Stanislava Tzaneva, Kornelia Böhler. "Comparative analysis of endovenous laser ablation versus ultrasound ‐ guided foam sclerotherapy for the treatment of venous leg ulcers", Dermatologic Therapy, 2022 Crossref	13 words — < 1%
8	H. Lu. "One-week regimens containing ranitidine bismuth citrate, furazolidone and either amoxicillin or tetracycline effectively eradicate Helicobacter pylori: a multicentre, randomized, double-blind study", Alimentary Pharmacology and Therapeutics, 12/2001 Crossref	12 words — < 1%
9	bmccancer.biomedcentral.com Internet	12 words — < 1%
10	www.science.gov Internet	12 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 12 WORDS