ISSN 2644-3236 (online)

Artificial Intelligence in *Gastroenterology*

Artif Intell Gastroenterol 2023 June 8; 4(1): 1-27

Published by Baishideng Publishing Group Inc

AIG

Artificial Intelligence in Gastroenterology

Contents

Quarterly Volume 4 Number 1 June 8, 2023

MINIREVIEWS

ORIGINAL ARTICLE

Retrospective Study

Artificial Intelligence in Gastroenterology

Quarterly Volume 4 Number 1 June 8, 2023

Contents

ABOUT COVER

Editorial Board Member of *Artificial Intelligence in Gastroenterology*, Haseeb Ahmad Khan, PhD, Full Professor, Department of Biochemistry, King Saud University, Riyadh 11451, Saudi Arabia. khan_haseeb@yahoo.com

AIMS AND SCOPE

The primary aim of *Artificial Intelligence in Gastroenterology* (*AIG*, *Artif Intell Gastroenterol*) is to provide scholars and readers from various fields of artificial intelligence in gastroenterology with a platform to publish high-quality basic and clinical research articles and communicate their research findings online.

AIG mainly publishes articles reporting research results obtained in the field of artificial intelligence in gastroenterology and covering a wide range of topics, including artificial intelligence in gastrointestinal cancer, liver cancer, pancreatic cancer, hepatitis B, hepatitis C, nonalcoholic fatty liver disease, inflammatory bowel disease, irritable bowel syndrome, and *Helicobacter pylori* infection.

INDEXING/ABSTRACTING

The *AIG* is now abstracted and indexed in Reference Citation Analysis, China Science and Technology Journal Database.

RESPONSIBLE EDITORS FOR THIS ISSUE

NAME OF JOURNAL
Artificial Intelligence in Gastroenterology

ISSN
ISSN 2644-3236 (online)

LAUNCH DATE
July 28, 2020

FREQUENCY
Quarterly

EDITORS-IN-CHIEF
Rajvinder Singh, Ferruccio Bonino

EDITORIAL BOARD MEMBERS
https://www.wjgnet.com/2644-3236/editorialboard.htm

PUBLICATION DATE
June 8, 2023

INSTRUCTIONS TO AUTHORS
https://www.wjgnet.com/bpg/gerinfo/204

GUIDELINES FOR ETHICS DOCUMENTS
https://www.wjgnet.com/bpg/GerInfo/287

GUIDELINES FOR NON-NATIVE SPEAKERS OF ENGLISH
https://www.wjgnet.com/bpg/gerinfo/240

PUBLICATION ETHICS
https://www.wjgnet.com/bpg/GerInfo/288

PUBLICATION MISCONDUCT
https://www.wjgnet.com/bpg/gerinfo/208

ARTICLE PROCESSING CHARGE
https://www.wjgnet.com/bpg/gerinfo/242

STEPS FOR SUBMITTING MANUSCRIPTS
https://www.wjgnet.com/bpg/GerInfo/239

ONLINE SUBMISSION
https://www.f6publishing.com

Artificial Intelligence in Gastroenterology

Submit a Manuscript: https://www.f6publishing.com

Artif Intell Gastroenterol 2023 June 8; 4(1): 1-9

DOI: 10.35712/aig.v4.i1.1

ISSN 2644-3236 (online)

MINIREVIEWS

Big data and variceal rebleeding prediction in cirrhosis patients

Quan Yuan, Wen-Long Zhao, Bo Qin

Specialty type: Gastroenterology and hepatology

Provenance and peer review: Unsolicited article; Externally peer reviewed.

Peer-review model: Single blind

Peer-review report's scientific quality classification

Grade A (Excellent): 0
Grade B (Very good): B
Grade C (Good): 0
Grade D (Fair): D
Grade E (Poor): 0

P-Reviewer: Byeon H, South Korea; Leowattana W, Thailand

Received: January 8, 2023
Peer-review started: January 8, 2023
First decision: January 21, 2023
Revised: February 3, 2023
Accepted: March 10, 2023
Article in press: March 10, 2023
Published online: June 8, 2023

Quan Yuan, Department of Gastroenterology, The First Affiliated Hospital of Chongqing Medical University, Chongqing 400042, China

Wen-Long Zhao, College of Medical Informatics, Chongqing Medical University, Chongqing 400016, China

Wen-Long Zhao, Medical Data Science Academy, Chongqing 400016, China

Wen-Long Zhao, Chongqing Engineering Research Centre for Clinical Big-data and Drug Evaluation, Chongqing 400016, China

Bo Qin, Department of Infectious Diseases, The First Affiliated Hospital of Chongqing Medical University, Chongqing 400042, China

Corresponding author: Bo Qin, MD, Professor, Department of Infectious Diseases, The First Affiliated Hospital of Chongqing Medical University, No. 1 Youyi Road, Yuzhong District, Chongqing 400042, China. qinbo@cqmu.edu.cn

Abstract

Big data has convincing merits in developing risk stratification strategies for diseases. The 6 "V"s of big data, namely, volume, velocity, variety, veracity, value, and variability, have shown promise for real-world scenarios. Big data can be applied to analyze health data and advance research in preclinical biology, medicine, and especially disease initiation, development, and control. A study design comprises data selection, inclusion and exclusion criteria, standard confirmation and cohort establishment, follow-up strategy, and events of interest. The development and efficiency verification of a prognosis model consists of deciding the data source, taking previous models as references while selecting candidate predictors, assessing model performance, choosing appropriate statistical methods, and model optimization. The model should be able to inform disease development and outcomes, such as predicting variceal rebleeding in patients with cirrhosis. Our work has merits beyond those of other colleagues with respect to cirrhosis patient screening and data source regarding variceal bleeding.

Key Words: Big data; Disease onset; Prognosis; Modeling; Cirrhosis; Gastrointestinal rebleeding

Core Tip: Big data have been applied in many fields including finance, traffic control, logistics, healthcare, and environmental protection. Modeling is an efficient method for completing various tasks, and verification of its validity is vital for ensuring high-quality operation and yielding satisfactory results. Predictor screening guarantees the establishment of a practical, convenient, and favorable model for prognosis prediction. Utilizing a regression model trained with numerous data mined from big data acquired from real-world hospitals is helpful for informing disease or status onset and its prognosis such as in variceal rebleeding, which is one of the leading causes of death in cirrhosis patients.

Citation: Yuan Q, Zhao WL, Qin B. Big data and variceal rebleeding prediction in cirrhosis patients. *Artif Intell Gastroenterol* 2023; 4(1): 1-9
URL: https://www.wjgnet.com/2644-3236/full/v4/i1/1.htm
DOI: https://dx.doi.org/10.35712/aig.v4.i1.1

INTRODUCTION

Many risk stratification strategies for diseases mainly depend on single-/medium-sized cohort studies or their meta-analysis[1,2], with lead-time bias taken into consideration[3,4]. This type of study method is, by design, well scheduled and well phenotyped but selective for the population sampled, which may not reflect the real-world, pan-subject profile. Real-world patients may have comorbidities, be taking concomitant medications, may be excluded from short-term follow-up, or have poor patient compliance. Direct data acquisition from basic healthcare institutions and cohorts is more representative than limited sampling.

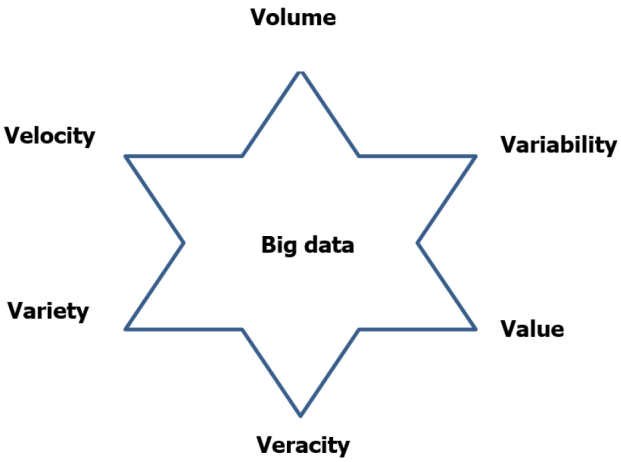
HISTORY OF BIG DATA

Although the use of piles of data in the medical field has a relatively long history[5-7], the term "big data" appeared only in the 1990s and quickly became popular[8-10]. "Big" is a relative term, especially when it relates to data. Big data usually refers to datasets that exceed the capabilities of commonly used software tools to store, manage, and process that amount of data within a suitable period of time[11]. The term is described by 315 characteristics[12] and fundamentally by the 6 "V"s: volume, velocity, variety, veracity, value, and variability[13-17] (Figure 1).

During the recent decade, methods for collecting, storing, and managing big data have evolved[18-20]. We are now entering an era of monitoring health changes using clinical indicators, such as vital signs, serum sugar, lipids, sweating, and bladder fullness, with wearable devices[11]. These changes can reflect physiological change. Constant variation and altered levels may result in different pathological states. Here, we review the applications of big data in predicting disease onset and prognosis, especially variceal rebleeding prediction in cirrhosis patients.

APPLICATIONS OF BIG DATA

Applications of big data include its use as a tool to monitor the onset of conditions and diseases. Big data have been used for this purpose in relation to hypertension[21], pediatric oncology[22], oral care [23], general practice[24], rheumatic diseases[25], renal diseases[26], mechanical ventilation management in the intensive care unit[27], and cirrhosis and hepatocellular carcinoma morbidity in the nonalcoholic fatty liver disease/nonalcoholic steatohepatitis population[28]. Situations such as the commencement, development, and control of diseases can be studied and visualized using big data techniques, which is a promising and beneficial approach. With the help of big data, the creation of large, collaborative data can lay a more solid foundation for robust data sharing and scientific discovery in predicting the onset of pediatric oncology. Registry-based research, however, is one of the conventional research methods regarding pediatric cancers. In these studies, a multisite registry for the study of pediatric patients was utilized, including fields of descriptive epidemiology, survivors, genomics, new registry description, data harmonization, palliative and supportive care, radiology, consensus guidelines, hereditary pediatric cancer, electronic health records, and prospective clinical trials. Limitations of registry-based research include the latest publication time range only, a restricted single publication database, and a limited amount of research and registries only if they have yielded publicly-published peer-reviewed papers[22]. With this study strategy, data cannot be automatically mined, cleaned, and integrated to perfect the already existing study. When it comes to new subjects, we need to redo the statistical analysis, while modeling and machine study in the big data scenario can perform the



DOI: 10.35712/aig.v4.i1.1

Figure 1 Six "V"s of big data.

whole analysis process.

Healthcare data in some regions are complete and accessible for analysis. Real-world data from primary healthcare facilities in communities in European countries are a good resource, as the primary healthcare service is state-covered and there are few or no co-payments. Therefore, healthcare information and data are collected and stored by state-run big data centers. Most residents are registered at birth and have their complete healthcare information in electronic form, which can be accessed by regional practitioners and analyzed for real-world application scenarios[29]. However, numerous parameters, especially administrative data, mined from patients' inpatient and outpatient Hospital Information System/Electronic Medical Record system *via* various algorithms are at risk of information and privacy leaking. Therefore, preliminary selection of data, especially low-dimensional administrative data, is preferable to decrease information leakage and privacy invasion.

Big data boosts the depth and breadth of research in fundamental biology and clinical medicine. There is already impressive progress due to this, including in exome sequencing[30], genomics, and proteomics. Taking the coronavirus disease 2019 pandemic as an example, primary research, clinical practices regarding treatment, and even trends in media campaigns of whether or not executing lockdown and a positive policy of nucleic acid testing can be swiftly analyzed with big data tools to assist epidemic control[31].

STUDY DESIGN

Study design comprises data source selection, inclusion and exclusion criteria, standard confirmation and cohort establishment, follow-up strategy, and events of interest. A multicountry European real-world study acquired patient data within a set research period mined from central transcription, laboratories, pharmacy offices, medical insurance departments, administrative departments, and other departmental databases *via* an electronic health record data repository along with molecular typing from molecular biology laboratories for preventing outbreaks of hospital infections[32]. Chart presentations can be used to analyze and interpret descriptive data. The Fib-4 score (age, aspartate aminotransferase, alanine aminotransferase, and platelets), which is composed of entirely non-invasive parameters, has been used to detect early liver fibrosis[28].

MODEL DEVELOPMENT AND EFFICACY VERIFICATION

With respect to development and efficiency verification of disease onset and prognosis models, researchers have performed extensive work. Model development is the process of collecting vital parameters (risk factors) of consequence and weighted with varied weight coefficients to form a weighted function. This requires the identification of predominant predictors from a large amount of preselected candidate predictors, assigning proper weights to each predictor to obtain a combined risk score, and assessing the model's predictive performance with statistical methods such as a calibration plot. The latter includes calibration, discrimination, and (re)classification properties, assessing its potential for generalization using internal validation techniques and if necessary optimizing the model to avoid overfitting. Data sources should preferably be prospective cohort(s) with a randomized controlled trial design or real-world medical record data. Preferred outcome choices are those that are

related to patients or individuals such as remission time and follow-up period. Methods for outcome verification should be included, and the blind method is preferred.

Regarding the selection of candidate predictors, a surplus should be defined and analyzed before finally including a subset in the final model. Incorporation bias should be avoided by blinding. Data quality control, missing data processing, continuous predictor modeling, final model development, relative weight assignment for each predictor, and internal validation are essential in the process of creating a final prediction model[33].

Choosing appropriate statistical methods during model establishment is vital to guarantee reliability and validity. Regression analysis, including univariate and multivariate regression, is the most commonly used statistical method, especially Cox regression[34] and LASSO[35]. The hazard ratio is used to differentiate cohorts across different conditions and coefficients. Featured with net benefit and threshold probability for more convenient yet trusty clinical decision making, decision curve analysis has been used to evaluate whether or not to use a certain prediction model[36]. In this approach, the theoretical relationship between the threshold probabilities of a disease (that a disease will take place) and the relative frequency of false positives and false negatives are examined to ensure the validity of a prediction model.

The benefits of applying decision curve analysis can be quantified as whether a model can be easily and effectively applied in clinical situations. Its ability to help compare several different models regarding one issue is another advantage[37]. The parameter indicating risk threshold "T value" has been used to study treatment decisions in risk models. The harm-to-benefit ratio is related to the T value, which is in line with the former. Balancing all benefits and harms in different scenarios is key to determining which T value is reasonable[38]. The net benefit (NB) value, which is a combined "net" effect of the true positives and false positives, was introduced to evaluate the potential clinical application of an estimating tool or a risk-predicting model. Setting the decisive threshold range in modeling is important, which is the boundary to determine whether a patient is judged as positive for a disease or not[39]. However, NB does not directly make up the harms and costs in acquiring the predictors for the chosen model. The focus of NB is to derive the best tradeoff between sufficient indicators and convenience in clinical application[40].

Model optimization should be conducted in order to reduce the number of predictors and avoid an unmanageable dataset or workload. AMSGrad ("far from the minimum"), a putative optimal method for optimizing models, is commonly used for low-cost cause. By switching to the direct linear method near the end of the optimization, AMSGrad can do its magic as it has long convergence tails[41]. As for multiobjective racing algorithms with fixed confidence, SPRINT-Race is the first algorithm developed and uses a nonparametric, ternary-decision, dual-sequential probability ratio test to infer a pairwise dominance or nondominance relationship. In order to minimize the computational effort, the probability of mistakenly erasing any Pareto-optimal models or returning any clearly dominating models is restricted, which can achieve a pre-estimated confidence level to ensure the quality of the models generated[42], by sequentially applying a Holm's step-down family-wise error rate control method. The quantification of model-to-data correspondence is pivotal to measure a model's performance and future application for the problem at hand. The *Drosophila melanogaster* gap gene system model demonstrated the importance of error quantification, and it is applicable to a wide array of developmental modeling studies[43]. The support vector machine, GLM-Net, generalized linear model, partial least squares, neural network, k-nearest neighbors, random forest, and boosted tree are useful tools for establishing the model to predict prognosis in patients with breast cancer[44]. Comparing their differences in performance and necessary model optimization can lead to better and more efficient application in practice.

PREDICTOR SCREENING FOR PROGNOSIS

Researchers have proposed methods for predictor screening with regard to disease prognosis, such as the Model for End-stage Live Disease (MELD) for cirrhosis-related mortality prediction and the APACHE model for critically ill patients. The clinical data of cirrhosis patients who had early admission, including clinical and socioeconomic factors, were mined from electronic medical records and classified for risk stratification in order to predict readmission within 30 d[45]. The European Organization for Research and Treatment of Cancer (EORTC) risk tables[46], which include six clinical and pathological factors (number of tumors, tumor size, prior recurrence rate, T category, carcinoma *in situ*, and grade), were recommended by the European Association of Urology and used to separately predict the short-term and long-term risks of progression and recurrence in an individual patient with a non-muscular invasive bladder tumor. It divided patients into four groups with individual recurrent and progression scores. However, as EORTC risk tables overestimated recurrence in all risk groups and progression in the high-risk group, the Club Urológico Español de Tratamiento Oncológico scoring model[47] was developed. The well-known new EORTC model[48], or European Association of Urology risk groups, was popular in recurrence and progression prediction, in which tumor diameter and extent were key predictors for progression prediction in multistate analyses. The health belief model has been

used for risk factors identifying aged Jordanian adults for prostate cancer screening[49]. Development and validation of a prediction model, including internal and external, temporal and geographical, domain validation, and their revision, are all crucial to identify predictors of prognosis[50].

RISK INDICATORS OF VARICEAL REBLEEDING IN CIRRHOSIS

Studies have reported several prediction models that predict variceal rebleeding in patients with cirrhosis. Risk indicators are components of prediction models. Invariably, studies in spotting possible risk indicators of variceal rebleeding among cirrhosis patients require a long study period. Child-Pugh score and hepatic-venous pressure gradient are the most significant prognostic factors in stratifying the probability of variceal rebleeding[51]. Antiviral treatment significantly reduced rebleeding in patients with hepatitis B virus (HBV)-related cirrhosis. In-time prophylactic endoscopic treatment of upper gastrointestinal varices after first-time bleeding, including endoscopic varix ligation (EVL) and gastric fundus varix gluing, is important in postponing variceal rebleeding[52]. Tachycardia, high creatinine level, and low albumin level are independent factors associated with rebleeding, suggesting a potential predictive role. The transverse of these variables into predictive scores may provide improved prognosis for patients with variceal bleeding[53]. Pre-emptive transjugular intrahepatic portosystemic shunt was independently related to a lower rebleeding rate[54]. Albumin transfusion in patients with low albumin levels was positively associated with a decreased rebleeding rate[55]. Five studies showed a lower rebleeding rate after EVL or drug therapy (non-selective β-blockers ± isosorbide mononitrate), and four trials found decreased variceal rebleeding with combined therapy (EVL+ non-selective β-blockers+ isosorbide mononitrate)[56].

However, some indicators have a negative function in preventing rebleeding. A multicenter, double-blind, parallel study of 158 patients indicated that taking simvastatin besides standard prophylaxis (rest, fluid restriction, preventing infection, regular endoscopic examination, anti-HBV therapy, non-selective β-blocker, *etc*.) did not decrease the rebleeding rate[57]. The rate of variceal rebleeding was not reduced after anticoagulation according to a single-center, prospective cohort study[58]. Worsened liver function or insensitive hemodynamic response to non-selective β-blockers indicated an elevated rebleeding rate [51]. A Chinese study of 3289 hospitalized patients who underwent EVL indicated that male sex, Child-Pugh score > 7.2, and volume of blood vomited before EVL were independent risk indicators of early rebleeding, while albumin concentration > 31.5 g/L was a protective indicator[59]. Bacterial infection in patients with variceal bleeding was strongly positively related to early rebleeding[60]. Acute-on-chronic liver failure is an independent risk factor of variceal rebleeding[54]. The presence of ascites or hepatic encephalopathy, MELD score > 12, or hepatic-venous pressure gradient > 20 mmHg indicated an elevated early (less than 6 wk) rebleeding rate[61].

The above indicators were then filtered and optimized by statistical methods, such as Cox regression or LASSO, and systemically integrated into a function with the help of programming or statistical software such as R, Python, SPSS, or SAS. This function was actually a preliminary prediction model.

SIGNIFICANCE OF PREDICTION MODELS

Models predicting disease onset and prognosis play an essential and sometimes surprising role as convenient assistants in planning prophylactic, therapeutic, and follow-up strategies. Traditionally, medical data such as medical history, results of physical examination, laboratory tests, imaging and endoscopic information, *etc*. were integrated by doctors' clinical comprehension or into patients' timelines drafted on a paper to identify how disease progressed and predicted the possible prognosis according to the trend in medical indicators. Prediction models free doctors from numerous medical data of patients with different diseases, complications, physical, psychological, and socioeconomic situations. All they need to do is to type prescribed parameters into the model and click! The results of the onset and prognosis of a given disease are then provided.

Prediction models are currently extensively applied in the medical field to inform individuals and healthcare providers on the risks of developing a particular disease, its outcome, and to guide doctors to make better decisions in mitigating these risks. A recent Chinese study indicated that the MELD score and MELD-Na score, including the R score, were useful in predicting variceal rebleeding[62]. Another study indicated that the MELD-Na score model, which indicates liver function, was more efficient than the MELD model and Child-Pugh score model in predicting rebleeding among cirrhosis patients who underwent EVL.

SAFETY AND PRIVACY CONCERNS

Last but not least, it is worth noting that models using low-dimensional administrative data outper-

formed in big data analysis with respect to decreasing information safety and privacy invasion. According to several studies, the models did not improve when high-resolution, privacy-invasive behavioral data were included[63]. De-ID software (De-ID Data) has been used to assign a study identification number to every enrolled patient. Therefore, criteria, included in the informed consent established by the research review board, for exemption from enrollment were met[32]. The *Drosophila melanogaster* gap gene system gives a good example of demonstrating the significance of error quantification, in which model parameters were optimized against *in situ* immunofluorescence intensities. It can be applied to other studies in various fields with regard to model development.

DISCUSSION

Gastrointestinal (GI) rebleeding is a leading cause of mortality in patients with cirrhosis, as massive GI bleeding can induce hemorrhagic shock, disseminated intravascular coagulation, and opportunistic infections, especially pulmonary infection and spontaneous bacterial peritonitis. Thus, reducing or postponing GI rebleeding is significant. A handy tool for clinicians that can be operated on smart phones or other mobile intelligent devices within seconds to evaluate the GI rebleeding rate is interesting and useful for risk grading. Just type in several common laboratory test indicators, click on "go," and the rebleeding rate and prognosis of a specific patient are provided.

Our work has merits beyond those of other colleagues. According to our literature retrieval on PubMed, there are no other studies on the prediction and prognosis analysis of GI rebleeding except for one article published last year indicating that the degree of liver stiffness is consistent with GI rebleeding rate in cirrhosis patients[64]. However, the above mentioned exclusive study has limitations. First, it was a prospective cohort study with only 289 patients enrolled in the final analysis, although PASS 15 was applied to calculate the statistically minimum sample size. In our ongoing study applying big data platform to evaluation the rebleeding rate of cirrhosis patients, we obtained real-world data from a big data platform collecting many more indicators from six hospitals, which were automatically collected. Second, our study included patients with esophageal and gastric fundus varices rebleeding, which were the most common varices presented in cirrhosis patients, and the other study only included esophageal varix rebleeding. Finally, the previous study only included patients with HBV-related decompensated cirrhosis, while our data were collected from cirrhosis patients with alcohol-related cirrhosis, autoimmune-related cirrhosis, primary biliary cirrhosis, and lipogenic cirrhosis in addition to HBV-related cirrhosis. Following parameter filtering and modeling, our study used a visual nomogram to demonstrate correlations among risk indicators, occurrence, and prognosis of GI rebleeding, which provides clinicians with a more explicit demonstration of all indicators and their effects on one page to easily and rapidly evaluate a patient to establish a strategy for further management and follow-up.

CONCLUSION

Modeling is popular using regression analysis and has vast applications in predicting disease occurrence and prognosis. However, modeling and its validation are not the ultimate objective in terms of healthcare provider's clinical participation and patients' health outcomes. They need to be applied and provide convenience for clinical practice. Studies on the application and optimization of these models should be designed and conducted, focusing on the utilization of existing and updated models and their impact on behavior and (self-) management of physicians, healthcare providers, and general individuals[65,66], especially in patients with decompensated cirrhosis at high risk of variceal rebleeding and mortality. For diagnostic and prognostic modeling with higher consistency and efficiency in predicting, treating, and following up decompensated cirrhosis, more comprehensive data and a clearer display mode are needed.

FOOTNOTES

Author contributions: Yuan Q selected the topic and performed the majority of conception, writing, and revision of the manuscript; Zhao WL provided think tank, platform with regard to big data, site for academic discussion, and revision suggestions for the manuscript; Qin B provided administrative help and was the instigator and coordinator of the study; All authors have read and approved the final manuscript.

Conflict-of-interest statement: All the authors declare that they have no conflicts of interest.

Country/Territory of origin: China

ORCID number: Quan Yuan 0000-0001-7761-4113; Bo Qin 0000-0002-7802-2854.

S-Editor: Liu JH
L-Editor: Filipodia
P-Editor: Liu JH

REFERENCES

1 **Koehler EM**, Schouten JN, Hansen BE, van Rooij FJ, Hofman A, Stricker BH, Janssen HL. Prevalence and risk factors of non-alcoholic fatty liver disease in the elderly: results from the Rotterdam study. *J Hepatol* 2012; **57**: 1305-1311 [PMID: 22871499 DOI: 10.1016/j.jhep.2012.07.028]
2 **Younossi ZM**, Koenig AB, Abdelatif D, Fazel Y, Henry L, Wymer M. Global epidemiology of nonalcoholic fatty liver disease-Meta-analytic assessment of prevalence, incidence, and outcomes. *Hepatology* 2016; **64**: 73-84 [PMID: 26707365 DOI: 10.1002/hep.28431]
3 **Facciorusso A**, Ferrusquía J, Muscatiello N. Lead time bias in estimating survival outcomes. *Gut* 2016; **65**: 538-539 [PMID: 26163490 DOI: 10.1136/gutjnl-2015-310199]
4 **Jansen RJ**, Alexander BH, Anderson KE, Church TR. Quantifying lead-time bias in risk factor studies of cancer through simulation. *Ann Epidemiol* 2013; **23**: 735-741 [PMID: 23988688 DOI: 10.1016/j.annepidem.2013.07.021]
5 **Graunt J**. Mathematical Demography. Berlin, Heidelberg: Springer, 1975: 11-20
6 **Dumbill E**. A Revolution That Will Transform How We Live, Work, and Think: An Interview with the Authors of Big Data. *Big Data* 2013; **1**: 73-77 [PMID: 27442060 DOI: 10.1089/big.2013.0016]
7 **Rothman KJ**. Lessons from John Graunt. *Lancet* 1996; **347**: 37-39 [PMID: 8531550 DOI: 10.1016/S0140-6736(96)91376-8]
8 **de Mauro A**, Greco M, Grimaldi M. A formal definition of big data based on its essential features. Lib Rev. 2016 Apr 4; 65: 122-135 [DOI: 10.1108/LR-06-2015-0061]
9 **John R**. Mashey. Big Data and the Next Wave of Infra Stress. USENIX: The Advanced Computing Systems Association. 1998. Available from: https://www.usenix.org/Legacy/event/usenix99/invited_talks/mashey.pdf
10 **Lohr S**. The Origins of 'Big Data': An Etymological Detective Story. The New York Times. B4. 2013. Available from: https://www.mendeley.com/catalogue/45aafddc-f02a-37bd-b201-c5edbcd31e82/
11 **Mirchev M**, Mircheva I, Kerekovska A. The Academic Viewpoint on Patient Data Ownership in the Context of Big Data: Scoping Review. *J Med Internet Res* 2020; **22**: e22214 [PMID: 32808934 DOI: 10.2196/22214]
12 **Kapil G**, Agrawal A, Khan RA. A Study of Big Data Characteristics. International Conference on Communication and Electronics Systems; ICCES'16; 2016 October 21-22, Coimbatore, India [DOI: 10.1109/CESYS.2016.7889917]
13 **Nobanee H**. A Bibliometric Review of Big Data in Finance. *Big Data* 2021; **9**: 73-78 [PMID: 33861644 DOI: 10.1089/big.2021.29044.edi]
14 **Beyer MA**, Laney D. The Importance of 'Big Data': A Definition. Gartner Inc. June 21, 2021. Available from: https://www.gartner.com/en/documents/2057415/the-importance-of-big-data-a-definition
15 **Tseng IL**. Big data: related technologies, challenges and future prospects. Computing reviews, 2015, 56: 476-477. Available from: https://www.zhangqiaokeyan.com/academic-journal-foreign_other_thesis/020411385974.html
16 **Dobre C**, Xhafa F. Intelligent services for big data science. *Future Gener Comput Syst* 2014; **37**: 267-281 [DOI: 10.1016/j.future.2013.07.014]
17 **Owais SS**, Hussein NS. Extract five categories CPIVW from the 9V's characteristics of the big data. *Int J Adv Comput Sci Appl* 2016; **7**: 254-258 [DOI: 10.14569/IJACSA.2016.070337]
18 **O'Driscoll A**, Daugelaite J, Sleator RD. 'Big data', Hadoop and cloud computing in genomics. *J Biomed Inform* 2013; **46**: 774-781 [PMID: 23872175 DOI: 10.1016/j.jbi.2013.07.001]
19 **Costa FF**. Big data in biomedicine. *Drug Discov Today* 2014; **19**: 433-440 [PMID: 24183925 DOI: 10.1016/j.drudis.2013.10.012]
20 **Luo J**, Wu M, Gopukumar D, Zhao Y. Big Data Application in Biomedical Research and Health Care: A Literature Review. *Biomed Inform Insights* 2016; **8**: 1-10 [PMID: 26843812 DOI: 10.4137/BII.S31559]
21 **Okada M**. Big data and real-world data-based medicine in the management of hypertension. *Hypertens Res* 2021; **44**: 147-153 [PMID: 33250517 DOI: 10.1038/s41440-020-00580-3]
22 **Major A**, Cox SM, Volchenboum SL. Using big data in pediatric oncology: Current applications and future directions. *Semin Oncol* 2020; **47**: 56-64 [PMID: 32229032 DOI: 10.1053/j.seminoncol.2020.02.006]
23 **Finkelstein J**, Zhang F, Levitin SA, Cappelli D. Using big data to promote precision oral health in the context of a learning healthcare system. *J Public Health Dent* 2020; **80** Suppl 1: S43-S58 [PMID: 31905246 DOI: 10.1111/jphd.12354]
24 **Waschkau A**, Wilfling D, Steinhäuser J. Are big data analytics helpful in caring for multimorbid patients in general practice? *BMC Fam Pract* 2019; **20**: 37 [PMID: 30813904 DOI: 10.1186/s12875-019-0928-5]
25 **Manrique de Lara A**, Peláez-Ballestas I. Big data and data processing in rheumatology: bioethical perspectives. *Clin Rheumatol* 2020; **39**: 1007-1014 [PMID: 32062767 DOI: 10.1007/s10067-020-04969-w]
26 **Yang C**, Kong G, Wang L, Zhang L, Zhao MH. Big data in nephrology: Are we ready for the change? *Nephrology (Carlton)* 2019; **24**: 1097-1102 [PMID: 31314170 DOI: 10.1111/nep.13636]

27 **Smallwood CD**. Monitoring Big Data During Mechanical Ventilation in the ICU. *Respir Care* 2020; **65**: 894-910 [PMID: 32457178 DOI: 10.4187/respcare.07500]

28 **Alexander M**, Loomis AK, van der Lei J, Duarte-Salles T, Prieto-Alhambra D, Ansell D, Pasqua A, Lapi F, Rijnbeek P, Mosseveld M, Waterworth DM, Kendrick S, Sattar N, Alazawi W. Risks and clinical predictors of cirrhosis and hepatocellular carcinoma diagnoses in adults with diagnosed NAFLD: real-world study of 18 million patients in four European cohorts. *BMC Med* 2019; **17**: 95 [PMID: 31104631 DOI: 10.1186/s12916-019-1321-x]

29 **Kringos D**, Boerma W, Bourgueil Y, Cartier T, Dedeu T, Hasvold T, Hutchinson A, Lember M, Oleszczyk M, Rotar Pavlic D, Svab I, Tedeschi P, Wilm S, Wilson A, Windak A, Van der Zee J, Groenewegen P. The strength of primary care in Europe: an international comparative study. *Br J Gen Pract* 2013; **63**: e742-e750 [PMID: 24267857 DOI: 10.3399/bjgp13X674422]

30 **Suwinski P**, Ong C, Ling MHT, Poh YM, Khan AM, Ong HS. Advancing Personalized Medicine Through the Application of Whole Exome Sequencing and Big Data Analytics. *Front Genet* 2019; **10**: 49 [PMID: 30809243 DOI: 10.3389/fgene.2019.00049]

31 **Jung JH**, Shin JI. Big Data Analysis of Media Reports Related to COVID-19. *Int J Environ Res Public Health* 2020; **17** [PMID: 32781727 DOI: 10.3390/ijerph17165688]

32 **Sundermann AJ**, Miller JK, Marsh JW, Saul MI, Shutt KA, Pacey M, Mustapha MM, Ayres A, Pasculle AW, Chen J, Snyder GM, Dubrawski AW, Harrison LH. Automated data mining of the electronic health record for investigation of healthcare-associated outbreaks. *Infect Control Hosp Epidemiol* 2019; **40**: 314-319 [PMID: 30773168 DOI: 10.1017/ice.2018.343]

33 **Moons KG**, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, Grobbee DE. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart* 2012; **98**: 683-690 [PMID: 22397945 DOI: 10.1136/heartjnl-2011-301246]

34 **In J**, Lee DK. Survival analysis: part II - applied clinical data analysis. *Korean J Anesthesiol* 2019; **72**: 441-457 [PMID: 31096731 DOI: 10.4097/kja.19183]

35 **Tibshirani R**. The lasso method for variable selection in the Cox model. *Stat Med* 1997; **16**: 385-395 [PMID: 9044528 DOI: 10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3]

36 **Van Calster B**, Wynants L, Verbeek JFM, Verbakel JY, Christodoulou E, Vickers AJ, Roobol MJ, Steyerberg EW. Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators. *Eur Urol* 2018; **74**: 796-804 [PMID: 30241973 DOI: 10.1016/j.eururo.2018.08.038]

37 **Vickers AJ**, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006; **26**: 565-574 [PMID: 17099194 DOI: 10.1177/0272989X06295361]

38 **Pauker SG**, Kassirer JP. Therapeutic decision making: a cost-benefit analysis. *N Engl J Med* 1975; **293**: 229-234 [PMID: 1143303 DOI: 10.1056/NEJM197507312930505]

39 **Vickers AJ**, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016; **352**: i6 [PMID: 26810254 DOI: 10.1136/bmj.i6]

40 **Baker SG**, Kramer BS. Evaluating Prognostic Markers Using Relative Utility Curves and Test Tradeoffs. *J Clin Oncol* 2015; **33**: 2578-2580 [PMID: 26124476 DOI: 10.1200/JCO.2014.58.0092]

41 **Sabzevari I**, Mahajan A, Sharma S. An accelerated linear method for optimizing non-linear wavefunctions in variational Monte Carlo. *J Chem Phys* 2020; **152**: 024111 [PMID: 31941334 DOI: 10.1063/1.5125803]

42 **Zhang T**, Georgiopoulos M, Anagnostopoulos GC. Pareto-Optimal Model Selection *via* SPRINT-Race. *IEEE Trans Cybern* 2018; **48**: 596-610 [PMID: 28166512 DOI: 10.1109/TCYB.2017.2647821]

43 **Hengenius JB**, Gribskov M, Rundell AE, Umulis DM. Making models match measurements: model optimization for morphogen patterning networks. *Semin Cell Dev Biol* 2014; **35**: 109-123 [PMID: 25016297 DOI: 10.1016/j.semcdb.2014.06.017]

44 **Boughorbel S**, Al-Ali R, Elkum N. Model Comparison for Breast Cancer Prognosis Based on Clinical Data. *PLoS One* 2016; **11**: e0146413 [PMID: 26771838 DOI: 10.1371/journal.pone.0146413]

45 **Singal AG**, Rahimi RS, Clark C, Ma Y, Cuthbert JA, Rockey DC, Amarasingham R. An automated model using electronic medical record data identifies patients with cirrhosis at high risk for readmission. *Clin Gastroenterol Hepatol* 2013; **11**: 1335-1341.e1 [PMID: 23591286 DOI: 10.1016/j.cgh.2013.03.022]

46 **Sylvester RJ**, van der Meijden AP, Oosterlinck W, Witjes JA, Bouffioux C, Denis L, Newling DW, Kurth K. Predicting recurrence and progression in individual patients with stage Ta T1 bladder cancer using EORTC risk tables: a combined analysis of 2596 patients from seven EORTC trials. *Eur Urol* 2006; **49**: 466-5; discussion 475 [PMID: 16442208 DOI: 10.1016/j.eururo.2005.12.031]

47 **Fernandez-Gomez J**, Madero R, Solsona E, Unda M, Martinez-Piñeiro L, Gonzalez M, Portillo J, Ojea A, Pertusa C, Rodriguez-Molina J, Camacho JE, Rabadan M, Astobieta A, Montesinos M, Isorna S, Muntañola P, Gimeno A, Blas M, Martinez-Piñeiro JA. Predicting nonmuscle invasive bladder cancer recurrence and progression in patients treated with bacillus Calmette-Guerin: the CUETO scoring model. *J Urol* 2009; **182**: 2195-2203 [PMID: 19758621 DOI: 10.1016/j.juro.2009.07.016]

48 **Cambier S**, Sylvester RJ, Collette L, Gontero P, Brausi MA, van Andel G, Kirkels WJ, Silva FC, Oosterlinck W, Prescott S, Kirkali Z, Powell PH, de Reijke TM, Turkeri L, Collette S, Oddens J. EORTC Nomograms and Risk Groups for Predicting Recurrence, Progression, and Disease-specific and Overall Survival in Non-Muscle-invasive Stage Ta-T1 Urothelial Bladder Cancer Patients Treated with 1-3 Years of Maintenance Bacillus Calmette-Guérin. *Eur Urol* 2016; **69**: 60-69 [PMID: 26210894 DOI: 10.1016/j.eururo.2015.06.045]

49 **Abuadas MH**, Petro-Nustas W, Albikawi ZF. Predictors of Participation in Prostate Cancer Screening among Older Men in Jordan. *Asian Pac J Cancer Prev* 2015; **16**: 5377-5383 [PMID: 26225681 DOI: 10.7314/apjcp.2015.16.13.5377]

50 **Moons KG**, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, Woodward M. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012; **98**: 691-698 [PMID: 22397946 DOI: 10.1136/heartjnl-2011-301247]

51 **Magaz M**, Baiges A, Hernández-Gea V. Precision medicine in variceal bleeding: Are we there yet? *J Hepatol* 2020; **72**:

774-784 [PMID: 31981725 DOI: 10.1016/j.jhep.2020.01.008]

52 **He L**, Ye X, Ma J, Li P, Jiang Y, Hu J, Yang J, Zhou Y, Liang X, Lin Y, Wei H. Antiviral therapy reduces rebleeding rate in patients with hepatitis B-related cirrhosis with acute variceal bleeding after endotherapy. *BMC Gastroenterol* 2019; **19**: 101 [PMID: 31226942 DOI: 10.1186/s12876-019-1020-2]

53 **Jiménez Rosales R**, Martínez-Cara JG, Vadillo-Calles F, Ortega-Suazo EJ, Abellán-Alfocea P, Redondo-Cerezo E. Analysis of rebleeding in cases of an upper gastrointestinal bleed in a single center series. *Rev Esp Enferm Dig* 2019; **111**: 189-192 [PMID: 30569727 DOI: 10.17235/reed.2018.5702/2018]

54 **Trebicka J**, Gu W, Ibáñez-Samaniego L, Hernández-Gea V, Pitarch C, Garcia E, Procopet B, Giráldez Á, Amitrano L, Villanueva C, Thabut D, Silva-Junior G, Martinez J, Genescà J, Bureau C, Llop E, Laleman W, Palazon JM, Castellote J, Rodrigues S, Gluud L, Ferreira CN, Barcelo R, Cañete N, Rodríguez M, Ferlitsch A, Mundi JL, Gronbaek H, Hernández-Guerra M, Sassatelli R, Dell'Era A, Senzolo M, Abraldes JG, Romero-Gómez M, Zipprich A, Casas M, Masnou H, Primignani M, Weiss E, Catalina MV, Erasmus HP, Uschner FE, Schulz M, Brol MJ, Praktiknjo M, Chang J, Krag A, Nevens F, Calleja JL, Robic MA, Conejo I, Albillos A, Rudler M, Alvarado E, Guardascione MA, Tantau M, Bosch J, Torres F, Pavesi M, Garcia-Pagán JC, Jansen C, Bañares R; International Variceal Bleeding Observational Study Group and Baveno Cooperation. Rebleeding and mortality risk are increased by ACLF but reduced by pre-emptive TIPS. *J Hepatol* 2020; **73**: 1082-1091 [PMID: 32339602 DOI: 10.1016/j.jhep.2020.04.024]

55 **Wang Z**, Xie YW, Lu Q, Yan HL, Liu XB, Long Y, Zhang X, Yang JL. The impact of albumin infusion on the risk of rebleeding and in-hospital mortality in cirrhotic patients admitted for acute gastrointestinal bleeding: a retrospective study of a single institute. *BMC Gastroenterol* 2020; **20**: 198 [PMID: 32576140 DOI: 10.1186/s12876-020-01337-5]

56 **Puente A**, Hernández-Gea V, Graupera I, Roque M, Colomo A, Poca M, Aracil C, Gich I, Guarner C, Villanueva C. Drugs plus ligation to prevent rebleeding in cirrhosis: an updated systematic review. *Liver Int* 2014; **34**: 823-833 [PMID: 24373180 DOI: 10.1111/liv.12452]

57 **Abraldes JG**, Villanueva C, Aracil C, Turnes J, Hernandez-Guerra M, Genesca J, Rodriguez M, Castellote J, García-Pagán JC, Torres F, Calleja JL, Albillos A, Bosch J; BLEPS Study Group. Addition of Simvastatin to Standard Therapy for the Prevention of Variceal Rebleeding Does Not Reduce Rebleeding but Increases Survival in Patients With Cirrhosis. *Gastroenterology* 2016; **150**: 1160-1170.e3 [PMID: 26774179 DOI: 10.1053/j.gastro.2016.01.004]

58 **Amitrano L**, Guardascione MA, Scaglione M, Menchise A, Martino R, Manguso F, Lanza AG, Lampasi F. Splanchnic vein thrombosis and variceal rebleeding in patients with cirrhosis. *Eur J Gastroenterol Hepatol* 2012; **24**: 1381-1385 [PMID: 23114742 DOI: 10.1097/MEG.0b013e328357d5d4]

59 **Zhou JN**, Wei Z, Sun ZQ. [Risk factors for early rebleeding after esophageal variceal ligation in patients with liver cirrhosis]. *Zhonghua Gan Zang Bing Za Zhi* 2016; **24**: 486-492 [PMID: 27784425 DOI: 10.3760/cma.j.issn.1007-3418.2016.07.002]

60 **Boursier J**, Asfar P, Joly-Guillou ML, Calès P. [Infection and variceal bleeding in cirrhosis]. *Gastroenterol Clin Biol* 2007; **31**: 27-38 [PMID: 17273129 DOI: 10.1016/s0399-8320(07)89324-9]

61 **Ardevol A**, Alvarado-Tapias E, Garcia-Guix M, Brujats A, Gonzalez L, Hernández-Gea V, Aracil C, Pavel O, Cuyas B, Graupera I, Colomo A, Poca M, Torras X, Concepción M, Villanueva C. Early rebleeding increases mortality of variecal bleeders on secondary prophylaxis with β-blockers and ligation. *Dig Liver Dis* 2020; **52**: 1017-1025 [PMID: 32653417 DOI: 10.1016/j.dld.2020.06.005]

62 **Ma JL**, Chen X, He LL, Wei HS, Li P. Predictive value of child Pugh score, MELD score, MELD-Na score, APASAL score and R score in rebleeding and death of liver cirrhosis with esophagogastric varices. JCTH 2020; 36: 1278-1283. Available from: https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CJFD&dbname=CJFDLAST2020&filename=LCGD202006022&uniplatform=NZKPT&v=oJBqDh7hJVjoGmr584nNKS-V3t9sAkvjdOfHnm_2cgW89jdFm-acWVMwM3_eZulC

63 **Bjerre-Nielsen A**, Kassarnig V, Lassen DD, Lehmann S. Task-specific information outperforms surveillance-style big data in predictive analytics. *Proc Natl Acad Sci USA* 2021; **118** [PMID: 33790010 DOI: 10.1073/pnas.2020258118]

64 **Liu L**, Liu Q, Xiao N, Zhang Y, Nie Y, Zhu X. A Liver Stiffness Measurement-Based Nomogram Predicts Variceal Rebleeding in Hepatitis B-Related Cirrhosis. *Dis Markers* 2022; **2022**: 4107877 [PMID: 35692881 DOI: 10.1155/2022/4107877]

65 **Sourabh D**. Clinical Epidemiology: Principles, Methods and Applications for Clinical Research. D E Grobbee and A W Hoes. *Int J Epidemiol* 2010; **39**: 318-319 [DOI: 10.1093/ije/dyn349]

66 **Reilly BM**, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Ann Intern Med* 2006; **144**: 201-209 [PMID: 16461965 DOI: 10.7326/0003-4819-144-3-200602070-00009]

Artificial Intelligence in Gastroenterology

Submit a Manuscript: https://www.f6publishing.com

Artif Intell Gastroenterol 2023 June 8; 4(1): 10-27

DOI: 10.35712/aig.v4.i1.10

ISSN 2644-3236 (online)

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

Specialty type: Infectious diseases

Provenance and peer review: Invited article; Externally peer reviewed.

Peer-review model: Single blind

Peer-review report's scientific quality classification

Grade A (Excellent): 0
Grade B (Very good): 0
Grade C (Good): C
Grade D (Fair): D, D
Grade E (Poor): 0

P-Reviewer: Cui W, China; de Melo FF, Brazil; Kawabata H, Japan

Received: December 14, 2022
Peer-review started: December 14, 2022
First decision: January 22, 2023
Revised: April 1, 2023
Accepted: April 20, 2023
Article in press: April 20, 2023
Published online: June 8, 2023

Shahid Aziz, Tanvir Ahmad, Faisal Rasheed, Patients Diagnostic Lab, Isotope Application Division, Pakistan Institute of Nuclear Science and Technology, Islamabad 44000, Pakistan

Shahid Aziz, Rabaab Zahra, Department of Microbiology, Quaid-i-Azam University, Islamabad 45320, Pakistan

Shahid Aziz, Simone König, Interdisciplinary Center for Clinical Research, Core Unit Proteomics, University of Münster, Münster 48149, Germany

Muhammad Umer, Management Information System Division, Pakistan Institute of Nuclear Science and Technology, Islamabad 44000, Pakistan

Tayyab Saeed Akhter, Shafqat Iqbal, Centre for Liver and Digestive Diseases, Holy Family Hospital, Rawalpindi 46300, Pakistan

Maryum Ibrar, Pakistan Scientific and Technological Information Centre, Quaid-i-Azam University, Islamabad 45320, Pakistan

Tofeeq Ur-Rehman, Department of Pharmacy, Quaid-i-Azam University, Islamabad 45320, Pakistan

Alfizah Hanafiah, Faculty of Medicine, Department of Medical Microbiology and Immunology, Universiti Kebangsan Malaysia, Cheras, Kuala Lumpur 56000, Malaysia

Corresponding author: Shahid Aziz, PhD, Research Fellow, Patients Diagnostic Lab, Isotope Application Division, Pakistan Institute of Nuclear Science and Technology, Nilore, Islamabad 44000, Pakistan. saziz@bs.qau.edu.pk

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* (*H. pylori*) infection is a major risk factor.

AIM

To develop 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 diseases (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 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.

Key Words: Gastric cancer; Gastritis; Machine learning; Prediction model; *Helicobacter pylori*

Core Tip: This is a retrospective study to report the prevalence of *Helicobacter pylori* (*H. pylori*) infection in Pakistan along with its association with various risk factors having direct or indirect relationships with different gastroduodenal diseases (GDD) such as gastritis, ulcers, and gastric cancer (GC). GC risk factors were incorporated into a highly sensitive and specific dynamic computer tool for the prediction of GC with an impressive > 80% confidence. This GC prediction model is freely available and may be used to reduce unnecessary invasive procedures such as endoscopies. The research study assists the healthcare authorities in their understanding of the burden of GDD and GC, which is intertwined with *H. pylori* infection.

Citation: Aziz S, König S, Umer M, Akhter TS, Iqbal S, Ibrar M, Ur-Rehman T, Ahmad T, Hanafiah A, Zahra R, Rasheed F. Risk factor profiles for gastric cancer prediction with respect to *Helicobacter pylori*: A study of a tertiary care hospital in Pakistan. *Artif Intell Gastroenterol* 2023; 4(1): 10-27
URL: https://www.wjgnet.com/2644-3236/full/v4/i1/10.htm
DOI: https://dx.doi.org/10.35712/aig.v4.i1.10

INTRODUCTION

Gastric cancer (GC) is the fourth most common cancer in the world and the second-most common cause of cancer-related deaths[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* (*H. 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.

GC risk factors include age[11], gender[12] and all factors which are commonly named as general health risks such as smoking[13,14], alcohol and junk food consumption as well as reduced physical exercise[5,6,15,16]. Diet and, in particular, controlled sugar and salt consumption play a specific role in GC prevention[17-19]. Proton pump inhibitors (PPI), which are routinely prescribed in the management of gastric-acid-related disorders, may also pose a risk, when improperly used[20,21]. Harvard University adds in its "10 commandments of cancer prevention"[22] 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[23]. High chili consumption in some regions of South America, on the other hand, sensitizes the mucosa and poses a cancer risk[24].

Histological examination of gastric biopsies is currently the gold standard for GC diagnosis[15]. 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[25]. Guidelines were published in what instances endoscopic biopsies should be performed[26], 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[27]. 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[28].

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[29-31]. 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[32-34], most recently for the prediction of GC risk after *H. pylori* eradication[34]. 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[32], radiology[35], and the detection of the cancerous and precancerous lesions in the intestine[36].

MATERIALS AND METHODS

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 persistent 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 d which may effects 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 erythema or scanty erosions), moderate (neither mild nor marked), and marked (diffuse erythema, nodularity, hypertrophy of gastric folds and friability of gastric mucosa) according to Kyoto classification system [37]. 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[29]. 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[38] and those who did not[39] as well as studies, which evaluated the influence of red and processed meat[40], high salt consumption due to salted fish and meat[19], 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, PPI, non-steroidal anti-inflammatory drugs and other medicines was recorded. Categories of responses were defined as listed in Table 1[41].

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.

Nuclear stable isotope ^{13}C UBT: Active *H. pylori* infection was determined using non-invasive nuclear stable isotope ^{13}C UBT as described previously[29]. 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, United States) 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 spectrometer (Thermo Finnigan, Germany) and Delta V Plus mass spectrometer (Thermo Scientific, United States). A change in the δ ^{13}C value over baseline of more than 3‰ was considered positive.

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.

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.

HPE: Gastric (antrum and corpus) biopsy specimens were processed for histopathological examination according to the Operative Link for Gastritis Assessment (OLGA/OLGIM) scoring[42] alongside with Lauren and World Health Organization (WHO) classification systems[43] for the determination of NGM, gastritis, gastric ulcer, duodenal ulcer and GC differentiation and invasions.

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. Principal components analysis (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, United States).

Table 1 Score response categories of Likert scale questionnaire

Study variables/risk factors	Category	Consumption/behavior frequency in d/wk
Tooth brushing and miswak usage	Always	7
	Often	4-6
	Seldom	1-3
	Never	0
Consumption of chili, dairy products, rice, potatoes, red and processed meat, sweets, junk food	No	0
	Rarely	1-2
	Moderately	2-4
	Frequently	5-6
		Servings per day
Drinking black and green tea	Normal	1-2
	Moderate	3-4
	High	5-7
		Habits per day
Washing hands with soap before meal and after use of toilet	Always	7
	Often	4-6
	Seldom	1-3
	Never	0
Addiction and passive smoking	No	< 10 (In Pakistan smoking is common practice.)
	Yes	> 10
		Consumption of cooked food
Salt/sodium chloride consumption; Normal: 2300 mg/d; Low: 140 mg/serving; High: > 3400 mg/d[38]	No	Without salt (Patients with high blood pressure did not use salt in their food)
	Normal	Without adding salt
	Low	With additional pinch of salt/serving; 1 Pinch = 0.36 g or 360 mg
	High	After addition of several pinches of salt/serving

RESULTS

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 the Supplementary Excel file of Supplementary material. 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).

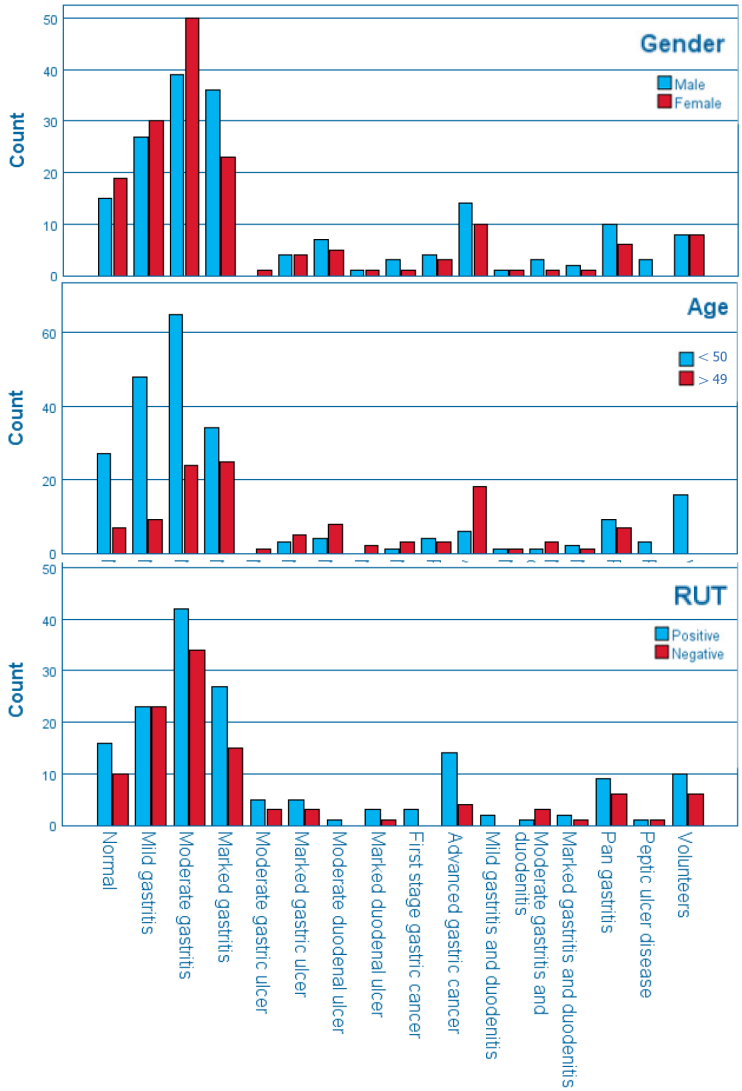
Descriptive characteristics of the cohort and results of the Chi-squared (χ^2) test to assess the association of socioeconomic demographics, risk factors, and histopathological grades among the three groups (NGM, GDD, GC) are presented in Table 2. Significant factors were age, education (one-third of 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 RUT results in Figure 1.

Table 2 Descriptive characteristics of the proband cohort groups and results of χ^2/Fisher's exact test (*P* value)

Factor	Categories	NGM	GDD	GC	*P* value	Significant
		% (number of patients/total number of patients)				
Infection status	Negative	28 (14/50)	39 (100/260)	26 (8/31)	0.176	> 0.05
	Positive	72 (36/50)	62 (160/260)	74 (23/31)		
Gender	Male	46 (23/50)	52 (136/260)	58 (18/31)	0.553	> 0.05
	Female	54 (27/50)	48 (124/260)	42 (13/31)		
BMI	Normal, underweight	58 (29/50)	63 (163/260)	78 (24/31)	0.191	> 0.05
	Overweight, obese	42 (21/50)	37 (97/260)	23 (7/31)		
Marital status	Married	78 (39/50)	80 (208/260)	87 (27/31)	0.580	> 0.05
	Single	22 (11/50)	20 (52/260)	13 (4/31)		
Age	< 46	82 (41/50)	65 (170/260)	29 (9/31)	**0.000**	**< 0.01**
	> 45	18 (9/50)	35 (90/260)	71 (22/31)		
Ethnic background	Federal	6 (3/50)	6 (16/260)	3 (1/31)	0.291[f]	> 0.05
	Lower punjab	16 (8/50)	15 (40/260)	13 (4/31)		
	Kashmir	12 (6/50)	5 (14/260)	3 (1/31)		
	Upper punjab	60 (30/50)	62 (161/260)	81 (25/31)		
	Khyber pakhtunkhwa	6 (3/50)	11 (29/260)	0 (0/31)		
Education level	Illiterate	18 (9/50)	35 (90/260)	48 (15/31)	0.013	**< 0.05**
	Literate	82 (41/50)	65 (170/260)	52 (16/31)		
Medication	Antibiotics	44 (22/50)	33 (86/260)	32 (10/31)	0.132[f]	> 0.05
	PPI	22 (11/50)	36 (93/260)	45 (14/31)		
	NSAID	12 (6/50)	4 (9/260)	3 (1/31)		
	Others	8 (4/50)	10 (27/260)	3 (1/31)		
	NIL	14 (7/50)	17 (45/260)	16 (5/31)		
Income level	10.000	6 (3/50)	7 (19/260)	3 (1/31)	0.007[f]	**< 0.05**
	11.000-30.000	48 (24/50)	69 (178/260)	84 (26/31)		
	> 30.000	46 (23/50)	24 (63/260)	13 (4/31)		
Acid reflux	No	66 (33/50)	45 (116/260)	39 (12/31)	0.013	**< 0.05**
	Yes	34 (17/50)	55 (144/260)	61 (19/31)		
Abdominal pain	No	54 (27/50)	25 (64/260)	10 (3/31)	0.000	**< 0.01**
	Yes	46 (23/50)	75 (196/260)	90 (28/31)		
Heartburn	No	56 (28/50)	46 (119/260)	45 (14/31)	0.403	> 0.05
	Yes	44 (22/50)	54 (141/260)	55 (17/31)		
Vomiting	No	64 (32/50)	41 (107/260)	16 (5/31)	0.000	**< 0.01**
	Yes	36 (18/50)	59 (153/260)	84 (26/31)		
Bloating	No	74 (37/50)	41 (106/260)	39 (12/31)	0.000	**< 0.01**
	Yes	26 (13/50)	59 (154/260)	61 (19/31)		
Black tea	Low	80 (40/50)	69 (178/260)	71 (22/31)	0.261	> 0.05
	High	20 (10/50)	32 (82/260)	29 (9/31)		
Green tea	Low	70 (35/50)	70 (181/260)	65 (20/31)	0.837	> 0.05
	High	30 (15/50)	30 (79/260)	36 (11/31)		
Chili consumption	Low	30 (15/50)	39 (102/260)	48 (15/31)	0.240	> 0.05

	High	70 (35/50)	61 (158/260)	52 (16/31)		
Dairy product consumption	Low	36 (18/50)	32 (83/260)	32 (10/31)	0.853	> 0.05
	High	64 (32/50)	68 (177/260)	68 (21/31)		
Fresh fruit & vegetable consumption	Low	28 (14/50)	18 (46/260)	23 (7/31)	0.222	> 0.05
	High	72 (36/50)	82 (214/260)	77 (24/31)		
Rice consumption	Low	26 (13/50)	37 (94/260)	23 (7/31)	0.154	> 0.05
	High	74 (37/50)	64 (166/260)	77 (24/31)		
Potato consumption	Low	32 (16/50)	40 (104/260)	32 (10/31)	0.422	> 0.05
	High	68 (34/50)	60 (156/260)	68 (21/31)		
Red meat consumption	Low	54 (27/50)	47 (123/260)	55 (17/31)	0.543	> 0.05
	High	46 (23/50)	53 (137/260)	45 (14/31)		
Processed meat consumption	Low	70 (35/50)	77 (199/260)	77 (24/31)	0.597	> 0.05
	High	30 (15/50)	24 (61/260)	23 (7/31)		
Junk food consumption	Low	78 (39/50)	84 (218/260)	84 (26/31)	0.596	> 0.05
	High	22 (11/50)	16 (42/260)	16 (5/31)		
Sweets consumption	Low	62 (31/50)	69 (178/260)	68 (21/31)	0.671	> 0.05
	High	38 (19/50)	32 (82/260)	32 (10/31)		
High salt intake	No	16 (8/50)	31 (80/260)	23 (7/31)	0.081	> 0.05
	Yes	84 (42/50)	69 (180/260)	77 (24/31)		
Overcrowding	Yes	54 (27/50)	42 (108/260)	55 (17/31)	0.129	> 0.05
	No	46 (23/50)	59 (152/260)	45 (14/31)		
Oral hygiene	Yes	84 (42/50)	87 (226/260)	77 (24/31)	0.340	> 0.05
	No	16 (8/50)	13 (34/260)	23 (7/31)		
Hand hygiene	Yes	98 (49/50)	99 (256/260)	97 (30/31)	0.437[f]	> 0.05
	No	2 (1/50)	2 (4/260)	3 (1/31)		
House insects	No	0 (0/50)	2 (4/260)	0 (0/31)	1.00[f]	> 0.05
	Yes	100 (50/50)	99 (256/260)	100 (31/31)		
Household animals	No	62 (31/50)	69 (180/260)	22/31 (71)	0.570	> 0.05
	Yes	38 (19/50)	31 (80/260)	29 (9/31)		
Potable water source	In	4 (2/50)	14 (37/260)	13 (4/31)	0.136	> 0.05
	Out	96 (48/50)	86 (223/260)	87 (27/31)		
Sewage system	Proper	0 (0/50)	3 (7/260)	7 (2/31)	0.196[f]	> 0.05
	Damaged	100 (50/50)	97 (253/260)	94 (29/31)		
Addiction	No	66 (33/50)	68 (176/260)	71 (22/31)	0.897	> 0.05
	Yes	34 (17/50)	32 (84/260)	29 (9/31)		
Passive smoking	No	58 (29/50)	62 (162/260)	65 (20/31)	0.806	> 0.05
	Yes	42 (21/50)	38 (98/260)	36 (11/31)		
Physical activity	Low	80 (40/50)	80 (208/260)	77 (24/31)	0.944	> 0.05
	High	20 (10/50)	20 (52/260)	23 (7/31)		
Family history of stomach disease	No	74 (37/50)	72 (187/260)	84 (26/31)	0.361	> 0.05
	Yes	26 (13/50)	28 (73/260)	16 (5/31)		
Type 2 diabetes	No	94 (47/50)	92 (238/260)	97 (30/31)	0.680[f]	> 0.05
	Yes	6 (3/50)	9 (22/260)	3 (1/31)		

[f]For the expected counts less than 5, *P* values were obtained from Fisher's exact test. Significant discriminators are marked in bold. BMI: Body mass index; NGM: Normal gastric mucosa, GC: Gastric cancer; GDD: Gastroduodenal diseases; NSAIDS: Non-steroidal anti-inflammatory drugs; PPI: Proton pump inhibitors.



DOI: 10.35712/aig.v4.i1.10

Figure 1 Cross-correlation bar charts for the study cohort with respect to gender, age and *Helicobacter pylori* infection status based on rapid urease test. RUT: Rapid urease test.

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 R^2) 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

Table 3 Factors associated with gastroduodenal diseases and gastric cancer *vs* normal gastric mucosa

Variable	Category	GDD/NGM		GC/GDD		GC/NGM	
		Significant	Odds ratio (95%CI)	Significant	Odds ratio (95%CI)	Significant	Odds ratio (95%CI)
Age	< 46	0.078	0.45 (0.18-1.09)	**0.00**	**0.19 (0.08-0.46)**	**0.000**	**0.08 (0.02-0.29)**
	> 45		Reference		Reference		Reference
Education level	Illiterate	0.404	1.44 (0.61-3.43)	0.609	1.25 (0.53-2.95)	0.325	1.81 (0.56-5.86)
	Literate		Reference		Reference		Reference
Income level	Low	0.767	1.25 (0.29-5.44)	0.847	0.79 (0.07-8.61)	0.992	0.99 (0.06-15.29)
	Middle	**0.028**	**2.32 (1.09-4.91)**	0.218	2.1 (0.65-6.8)	**0.022**	**4.86 (1.25-18.84)**
	Upper		Reference		Reference		Reference
Vomiting	No	0.088	0.54 (0.27-1.09)	0.599	1.26 (0.53-3.01)	**0.003**	**0.16 (0.05-0.53)**
	Yes		Reference		Reference		Reference
Bloating	No	**0.012**	**0.37 (0.17-0.8)**	**0.019**	**0.29 (0.1-0.81)**	0.184	0.47 (0.15-1.44)
	Yes		Reference		Reference		Reference
Abdominal pain	No	0.075	0.52 (0.25-1.07)	0.092	0.32 (0.09-1.2)	**0.016**	**0.17 (0.04-0.72)**
	Yes		Reference		Reference		Reference
Acid reflux	No	0.220	0.63 (0.3-1.32)	0.944	0.97 (0.41-2.3)	0.379	0.61 (0.2-1.83)
	Yes		Reference		Reference		Reference
Medication	Antibiotics	0.686	0.81 (0.29-2.25)	0.519	1.51 (0.43-5.24)	0.803	1.22 (0.25-5.85)
	PPI	0.520	1.45 (0.47-4.44)	0.139	2.48 (0.74-8.26)	0.118	3.58 (0.72-17.76)
	NSAIDS	**0.020**	**0.16 (0.03-0.75)**	0.936	1.1(0.1-11.93)	0.212	0.18 (0.01-2.68)
	Others	0.512	1.64 (0.38-7.12)	0.598	0.54 (0.06-5.3)	0.928	0.88 (0.06-12.75)
	Nil		Reference		Reference		Reference
High salt intake	No	0.215	1.77 (0.72-4.37)	0.215	0.55 (0.21-1.42)	0.965	0.97 (0.27-3.49)
	Yes		Reference		Reference		Reference

Nine variables were added to the multinomial logistic regression analysis with $P < 0.05$ (marked in bold). GDD: Gastroduodenal diseases; NGM: Normal gastric mucosa; NSAIDS: Non-steroidal anti-inflammatory drugs; PPI: Proton pump inhibitors; Sign.: Significant.

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, $P < 0.05$).

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, 9.1% (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.

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).

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 1). Decision tree analysis with a focus on GC (Supplementary Figure 1A) 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 (Supplementary Figure 1B), 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.

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 secondly, 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 patient's data.

The primary dataset (parameters in textual and structural format, Supplementary Excel file Training_Testing_Data of Supplementary material) 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 material). 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 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 (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.

Table 4 Correlation of histopathological variables, antral *vs* corpus biopsies ($P < 0.01$)

Variable	Spearman correlation coefficient
Helicobacter pylori load	0.991
Neutrophil infiltration	1.000
Mononuclear cell infiltration	0.942
Atrophy	0.969
Atypia	0.881
Metaplasia	1.000
Dysplasia	0.786
Atrophic score (OLGA)	0.951
Metaplasia score (OLGIM)	1.000
Inflammation	0.930
Gastritis	0.921
Ulceration	1.000
Eosinophilia	1.000

OLGA: Operative link for gastritis assessment; OLGIM: Operative link on gastric intestinal metaplasia.

Table 5 Gastric cancer associated risk factors chosen for prediction model building

Risk factors	Ref.
H. pylori infection	[1]
Family history	[4]
PPI	[8,20,21]
Addiction (smoking)	[13]
Passive smoking	[44,64]
Sewage system (cockroaches/*H. pylori*)	[31]
Potable water source (*H. pylori*)	[30]
Exercise/fruits and vegetables	[7,8]
BMI	[16]
Gender (male)	[12]
Age	[11]
High salt intake/green and black tea	[7,58]
Chili consumption	[24]
Processed food (meat) consumption	[19,40]
Sugar intake	[17]
Excess of rice and potatoes	[18]
Miswak usage	[23]

BMI: Body mass index; PPI: Proton pump inhibitors; *H. pylori*: *Helicobacter pylori*.

DISCUSSION

H. pylori infection is a serious public health problem with a high frequency among the population of developing countries[44]. 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%

Table 6 Dataset used to train the gastric cancer prediction model

Clinical findings	Total samples	Training dataset	Test dataset
NGM	50	35	15
Gastritis	232	162	70
Ulcer	30	21	9
GC	29	20	8
Total	**341**	**239**	**102**

GC: Gastric cancer; NGM: Normal gastric mucosa.

and 10%-30%, respectively[45]. Our previous study showed more than 70% frequency of *H. pylori* infection in the northern region of Pakistan[46]. Six years later[47], 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[10,48], while, here, 9.1% were calculated. In agreement with our previous findings[47], 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[45,49].

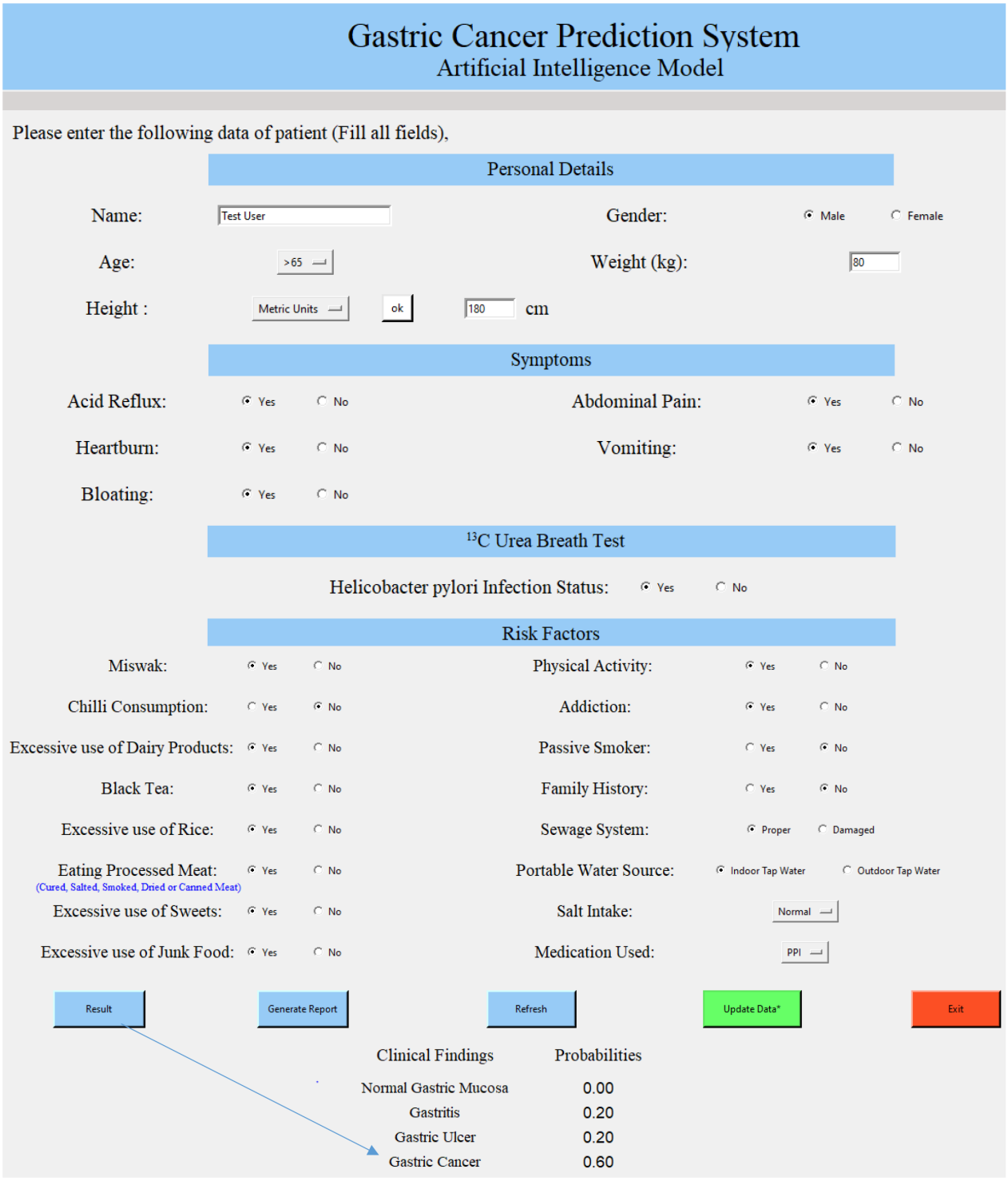
The increased risk of *H. pylori* positivity in developing countries has been associated with several environmental factors including lower socioeconomic conditions such as crowded households and poor hygiene[50]. Already in our previous study[29], these risk factors, further including pets and other 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 (51-139 USD; 11000-30000 PKR, 1 USD=215 PKR) where living conditions are difficult. About 256465 PKR (1194 USD) are required for appropriate living conditions and fulfillment of basic needs[51].

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[52]. Miswak has been traditionally used in Pakistan for oral hygiene due its antibacterial properties against both Gram positive and negative bacteria[23]. 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[53]. A study conducted in Korea indicated that high salt intake was associated with a higher risk of atrophic gastritis and intestinal metaplasia[54] and other authors showed that it can lead to the onset of pre-malignant lesions [55]. In addition, the carcinogenic effects of major *H. pylori* virulence factor cytotoxin associated gene A (*cagA*)-positive strains were increased[56,57]. We confirmed the higher risk of *H. pylori* infection (73%) in patients with a higher salt intake than 5 g/d as recommended by the WHO[58].

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[38]. 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[39] 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[39]. 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.



DOI: 10.35712/aig.v4.i1.10

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

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[50]. 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[49]. 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[32-34], 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[53]. 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* (*H. 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; and (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 urea breath test, rapid urease test, and histopathological examinations. A Random Forest (RF) -gastric cancer (GC) prediction model was 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. 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 be helpful in diagnosing other gastric diseases such as gastritis and ulcer and 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, Islamabad, 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.

FOOTNOTES

Author contributions: Rasheed F and Aziz S contributed to conceptualization; Aziz S contributed to methodology; Umer M contributed to software; Aziz S contributed to validation; Aziz S, König S and Ibrar M contributed to formal analysis; Rasheed F contributed to resources; Akhter ST and Iqbal S contributed to endoscopic procedures; Aziz S and König S contributed to writing – original draft preparation; Aziz S, König S, Ahmad T, Rasheed F, Hanafia A, and Rehman UT contributed to writing – review & editing; König S contributed to visualization; Zahra R and Rasheed F contributed to supervision; Aziz S and Rasheed F contributed to project administration; Aziz S and Rasheed F contributed to funding acquisition.

Institutional review board statement: 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).

Informed consent statement: The investigators explain the study to each patient and informed written consent was obtained to participate in this research and their clinical data was collected during interview using a questionnaire after endoscopic evaluation. Moreover, patients were also required to give informed consent to the study for analysis and publication of their anonymous clinical data.

Conflict-of-interest statement: All the authors declare no conflict of interest.

Data sharing statement: All the data has been shared in supplementary files.

Country/Territory of origin: Pakistan

ORCID number: Shahid Aziz 0000-0003-4388-0207; Simone König 0000-0003-0672-7246; Tayyab Saeed Akhter 0000-0002-7914-4626; Maryum Ibrar 0000-0002-0032-4959; Tofeeq Ur-Rehman 0000-0003-4904-1367; Rabaab Zahra 0000-0001-8784-0114; Faisal Rasheed 0000-0002-7703-9443.

Corresponding Author's Membership in Professional Societies: American Society for Microbiology, No. 200327988I.

S-Editor: Liu JH
L-Editor: A
P-Editor: Zhao S

REFERENCES

1 **Sitarz R**, Skierucha M, Mielko J, Offerhaus GJA, Maciejewski R, Polkowski WP. Gastric cancer: epidemiology, prevention, classification, and treatment. *Cancer Manag Res* 2018; **10**: 239-248 [PMID: 29445300 DOI: 10.2147/CMAR.S149619]

2 **Ferlay J**, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* 2015; **136**: E359-E386 [PMID: 25220842 DOI: 10.1002/ijc.29210]

3 Schistosomes, liver flukes and Helicobacter pylori. *IARC Monogr Eval Carcinog Risks Hum* 1994; **61**: 1-241 [PMID: 7715068]

4 **Yaghoobi M**, Bijarchi R, Narod SA. Family history and the risk of gastric cancer. *Br J Cancer* 2010; **102**: 237-242 [PMID: 19888225 DOI: 10.1038/sj.bjc.6605380]

5 **Wang Z**, Koh WP, Jin A, Wang R, Yuan JM. Composite protective lifestyle factors and risk of developing gastric adenocarcinoma: the Singapore Chinese Health Study. *Br J Cancer* 2017; **116**: 679-687 [PMID: 28125822 DOI: 10.1038/bjc.2017.7]

6 **Charvat H**, Sasazuki S, Inoue M, Iwasaki M, Sawada N, Shimazu T, Yamaji T, Tsugane S; JPHC Study Group. Prediction of the 10-year probability of gastric cancer occurrence in the Japanese population: the JPHC study cohort II. *Int J Cancer* 2016; **138**: 320-331 [PMID: 26219435 DOI: 10.1002/ijc.29705]

7 **Poorolajal J**, Moradi L, Mohammadi Y, Cheraghi Z, Gohari-Ensaf F. Risk factors for stomach cancer: a systematic review and meta-analysis. *Epidemiol Health* 2020; **42**: e2020004 [PMID: 32023777 DOI: 10.4178/epih.e2020004]

8 **Rawla P**, Barsouk A. Epidemiology of gastric cancer: global trends, risk factors and prevention. *Prz Gastroenterol* 2019; **14**: 26-38 [PMID: 30944675 DOI: 10.5114/pg.2018.80001]

9 **Daniyal M**, Ahmad S, Ahmad M, Asif HM, Akram M, Ur Rehman S, Sultana S. Risk Factors and Epidemiology of Gastric Cancer in Pakistan. *Asian Pac J Cancer Prev* 2015; **16**: 4821-4824 [PMID: 26163597 DOI: 10.7314/apjcp.2015.16.12.4821]

10 **Idrees R**, Fatima S, Abdul-Ghafar J, Raheem A, Ahmad Z. Cancer prevalence in Pakistan: meta-analysis of various published studies to determine variation in cancer figures resulting from marked population heterogeneity in different parts of the country. *World J Surg Oncol* 2018; **16**: 129 [PMID: 29976196 DOI: 10.1186/s12957-018-1429-z]

11 **Asaka M**, Kobayashi M, Kudo T, Akino K, Asaka Y, Fujimori K, Kikuchi S, Kawai S, Kato M. Gastric cancer deaths by age group in Japan: Outlook on preventive measures for elderly adults. *Cancer Sci* 2020; **111**: 3845-3853 [PMID: 32713120 DOI: 10.1111/cas.14586]

12 **Youn Nam S**, Park BJ, Nam JH, Ryu KH, Kook MC, Kim J, Lee WK. Association of current Helicobacter pylori infection and metabolic factors with gastric cancer in 35,519 subjects: A cross-sectional study. *United European Gastroenterol J* 2019; **7**: 287-296 [PMID: 31080613 DOI: 10.1177/2050640618819402]

13 **Praud D**, Rota M, Pelucchi C, Bertuccio P, Rosso T, Galeone C, Zhang ZF, Matsuo K, Ito H, Hu J, Johnson KC, Yu GP, Palli D, Ferraroni M, Muscat J, Lunet N, Peleteiro B, Malekzadeh R, Ye W, Song H, Zaridze D, Maximovitch D, Aragonés N, Castaño-Vinyals G, Vioque J, Navarrete-Muñoz EM, Pakseresht M, Pourfarzi F, Wolk A, Orsini N, Bellavia A, Håkansson N, Mu L, Pastorino R, Kurtz RC, Derakhshan MH, Lagiou A, Lagiou P, Boffetta P, Boccia S, Negri E, La Vecchia C. Cigarette smoking and gastric cancer in the Stomach Cancer Pooling (StoP) Project. *Eur J Cancer Prev* 2018; **27**: 124-133 [PMID: 27560662 DOI: 10.1097/CEJ.0000000000000290]

14 **Mao Y**, Hu J, Semenciw R, White K; Canadian Cancer Registries Epidemiology Research Group. Active and passive smoking and the risk of stomach cancer, by subsite, in Canada. *Eur J Cancer Prev* 2002; **11**: 27-38 [PMID: 11917206 DOI: 10.1097/00008469-200202000-00005]

15 **Cai Q**, Zhu C, Yuan Y, Feng Q, Feng Y, Hao Y, Li J, Zhang K, Ye G, Ye L, Lv N, Zhang S, Liu C, Li M, Liu Q, Li R, Pan J, Yang X, Zhu X, Li Y, Lao B, Ling A, Chen H, Li X, Xu P, Zhou J, Liu B, Du Z, Du Y, Li Z; Gastrointestinal Early Cancer Prevention & Treatment Alliance of China (GECA). Development and validation of a prediction rule for estimating gastric cancer risk in the Chinese high-risk population: a nationwide multicentre study. *Gut* 2019; **68**: 1576-1587 [PMID: 30926654 DOI: 10.1136/gutjnl-2018-317556]

16 **Wang J**, Yang DL, Chen ZZ, Gou BF. Associations of body mass index with cancer incidence among populations, genders, and menopausal status: A systematic review and meta-analysis. *Cancer Epidemiol* 2016; **42**: 1-8 [PMID: 26946037 DOI: 10.1016/j.canep.2016.02.010]

17 **Tay SW**, Li JW, Fock KM. Diet and cancer of the esophagus and stomach. *Curr Opin Gastroenterol* 2021; **37**: 158-163 [PMID: 33315794 DOI: 10.1097/MOG.0000000000000700]

18 **De Stefani E**, Correa P, Boffetta P, Deneo-Pellegrini H, Ronco AL, Mendilaharsu M. Dietary patterns and risk of gastric cancer: a case-control study in Uruguay. *Gastric Cancer* 2004; **7**: 211-220 [PMID: 15616769 DOI: 10.1007/s10120-004-0295-2]

19 **Lin SH**, Li YH, Leung K, Huang CY, Wang XR. Salt processed food and gastric cancer in a Chinese population. *Asian Pac J Cancer Prev* 2014; **15**: 5293-5298 [PMID: 25040991 DOI: 10.7314/apjcp.2014.15.13.5293]

20 **Naunton M**, Peterson GM, Bleasel MD. Overuse of proton pump inhibitors. *J Clin Pharm Ther* 2000; **25**: 333-340 [PMID: 11123484 DOI: 10.1046/j.1365-2710.2000.00312.x]

21 **Yibirin M**, De Oliveira D, Valera R, Plitt AE, Lutgen S. Adverse Effects Associated with Proton Pump Inhibitor Use. *Cureus* 2021; **13**: e12759 [PMID: 33614352 DOI: 10.7759/cureus.12759]

22 **Harvard Health Publishing**. The 10 commandments of cancer prevention. Accessed on 16 Aug 2021. 2019. Available from: www.health.harvard.edu/newsletter_article/the-10-commandments-of-cancer-prevention

23 **Abhary M**, Al-Hazmi AA. Antibacterial activity of Miswak (Salvadora persica L.) extracts on oral hygiene. *Journal of Taibah University for Science* 2016; **10**: 513-520 [DOI: 10.1016/j.jtusci.2015.09.007]

24 **Du Y**, Lv Y, Zha W, Hong X, Luo Q. Chili Consumption and Risk of Gastric Cancer: A Meta-Analysis. *Nutr Cancer* 2021; **73**: 45-54 [PMID: 32241189 DOI: 10.1080/01635581.2020.1733625]

25 **Manes G**, Saibeni S, Pellegrini L, Picascia D, Pace F, Schettino M, Bezzio C, de Nucci G, Hassan C, Repici A; The Fast-

Track Endoscopy Study Group. Improvement in appropriateness and diagnostic yield of fast-track endoscopy during the COVID-19 pandemic in Northern Italy. *Endoscopy* 2021; **53**: 162-165 [PMID: 32942316 DOI: 10.1055/a-1265-3315]

26 **Peixoto A**, Silva M, Pereira P, Macedo G. Biopsies in Gastrointestinal Endoscopy: When and How. *GE Port J Gastroenterol* 2016; **23**: 19-27 [PMID: 28868426 DOI: 10.1016/j.jpge.2015.07.004]

27 **Sheiko MA**, Feinstein JA, Capocelli KE, Kramer RE. The concordance of endoscopic and histologic findings of 1000 pediatric EGDs. *Gastrointest Endosc* 2015; **81**: 1385-1391 [PMID: 25440693 DOI: 10.1016/j.gie.2014.09.010]

28 **Panarelli NC**. Infectious diseases of the upper gastrointestinal tract. *Histopathology* 2021; **78**: 70-87 [PMID: 33382485 DOI: 10.1111/his.14243]

29 **Rasheed F**, Ahmad T, Bilal R. Prevalence and risk factors of Helicobacter pylori infection among Pakistani population. *Pak J Med Sci* 2012; **28**: 661-665

30 **Azevedo NF**, Vieira MJ, Keevil CW. Establishment of a continuous model system to study Helicobacter pylori survival in potable water biofilms. *Water Sci Technol* 2003; **47**: 155-160 [PMID: 12701922]

31 **Imamura S**, Kita M, Yamaoka Y, Yamamoto T, Ishimaru A, Konishi H, Wakabayashi N, Mitsufuji S, Okanoue T, Imanishi J. Vector potential of cockroaches for Helicobacter pylori infection. *Am J Gastroenterol* 2003; **98**: 1500-1503 [PMID: 12873569 DOI: 10.1111/j.1572-0241.2003.07516.x]

32 **Yang YJ**, Bang CS. Application of artificial intelligence in gastroenterology. *World J Gastroenterol* 2019; **25**: 1666-1683 [PMID: 31011253 DOI: 10.3748/wjg.v25.i14.1666]

33 **Niu PH**, Zhao LL, Wu HL, Zhao DB, Chen YT. Artificial intelligence in gastric cancer: Application and future perspectives. *World J Gastroenterol* 2020; **26**: 5408-5419 [PMID: 33024393 DOI: 10.3748/wjg.v26.i36.5408]

34 **Leung WK**, Cheung KS, Li B, Law SYK, Lui TKL. Applications of machine learning models in the prediction of gastric cancer risk in patients after Helicobacter pylori eradication. *Aliment Pharmacol Ther* 2021; **53**: 864-872 [PMID: 33486805 DOI: 10.1111/apt.16272]

35 **Liew C**. The future of radiology augmented with Artificial Intelligence: A strategy for success. *Eur J Radiol* 2018; **102**: 152-156 [PMID: 29685530 DOI: 10.1016/j.ejrad.2018.03.019]

36 **Ai L**, Tian H, Chen Z, Chen H, Xu J, Fang JY. Systematic evaluation of supervised classifiers for fecal microbiota-based prediction of colorectal cancer. *Oncotarget* 2017; **8**: 9546-9556 [PMID: 28061434 DOI: 10.18632/oncotarget.14488]

37 **Toyoshima O**, Nishizawa T, Koike K. Endoscopic Kyoto classification of Helicobacter pylori infection and gastric cancer risk diagnosis. *World J Gastroenterol* 2020; **26**: 466-477 [PMID: 32089624 DOI: 10.3748/wjg.v26.i5.466]

38 **Westerik N**, Reid G, Sybesma W, Kort R. The Probiotic Lactobacillus rhamnosus for Alleviation of Helicobacter pylori-Associated Gastric Pathology in East Africa. *Front Microbiol* 2018; **9**: 1873 [PMID: 30154777 DOI: 10.3389/fmicb.2018.01873]

39 **Wang S**, Zhou M, Ji A, Zhang D, He J. Milk/dairy products consumption and gastric cancer: an update meta-analysis of epidemiological studies. *Oncotarget* 2018; **9**: 7126-7135 [PMID: 29467955 DOI: 10.18632/oncotarget.23496]

40 **Kim SR**, Kim K, Lee SA, Kwon SO, Lee JK, Keum N, Park SM. Effect of Red, Processed, and White Meat Consumption on the Risk of Gastric Cancer: An Overall and Dose–Response Meta-Analysis. *Nutrients* 2019; **11** [PMID: 30979076 DOI: 10.3390/nu11040826]

41 U.S. Department of Agriculture and U.S. Department of Health and Human Services. Dietary Guidelines for Americans, 2020-2025. 9th Edition. December 2020, Accessed on April April 13, 2023. Available from: DietaryGuidelines.gov

42 **Yue H**, Shan L, Bin L. The significance of OLGA and OLGIM staging systems in the risk assessment of gastric cancer: a systematic review and meta-analysis. *Gastric Cancer* 2018; **21**: 579-587 [PMID: 29460004 DOI: 10.1007/s10120-018-0812-3]

43 **Berlth F**, Bollschweiler E, Drebber U, Hoelscher AH, Moenig S. Pathohistological classification systems in gastric cancer: diagnostic relevance and prognostic value. *World J Gastroenterol* 2014; **20**: 5679-5684 [PMID: 24914328 DOI: 10.3748/wjg.v20.i19.5679]

44 **Guevara B**, Cogdill AG. Helicobacter pylori: A Review of Current Diagnostic and Management Strategies. *Dig Dis Sci* 2020; **65**: 1917-1931 [PMID: 32170476 DOI: 10.1007/s10620-020-06193-7]

45 **Wang W**, Jiang W, Zhu S, Sun X, Li P, Liu K, Liu H, Gu J, Zhang S. Assessment of prevalence and risk factors of helicobacter pylori infection in an oilfield Community in Hebei, China. *BMC Gastroenterol* 2019; **19**: 186 [PMID: 31726980 DOI: 10.1186/s12876-019-1108-8]

46 **Rasheed F**, Ahmad T, Bilal R. Frequency of Helicobacter pylori infection using 13C-UBT in asymptomatic individuals of Barakaho, Islamabad, Pakistan. *J Coll Physicians Surg Pak* 2011; **21**: 379-381 [PMID: 21712001]

47 **Rasheed F**, Yameen A, Ahmad T, Bilal R. Rate of active Helicobacter pylori infection among symptomatic patients of Pakistan. *Malays J Pathol* 2017; **39**: 69-72 [PMID: 28413207]

48 **Rasheed F**, Campbell BJ, Alfizah H, Varro A, Zahra R, Yamaoka Y, Pritchard DM. Analysis of clinical isolates of Helicobacter pylori in Pakistan reveals high degrees of pathogenicity and high frequencies of antibiotic resistance. *Helicobacter* 2014; **19**: 387-399 [PMID: 24827414 DOI: 10.1111/hel.12142]

49 **Fareed R**, Abbas Z, Shah MA. Effect of Helicobacter pylori density on inflammatory activity in stomach. *J Pak Med Assoc* 2000; **50**: 148-151 [PMID: 11242713]

50 **Kouitcheu Mabeku LB**, Noundjeu Ngamga ML, Leundji H. Potential risk factors and prevalence of Helicobacter pylori infection among adult patients with dyspepsia symptoms in Cameroon. *BMC Infect Dis* 2018; **18**: 278 [PMID: 29907086 DOI: 10.1186/s12879-018-3146-1]

51 **Ceicdata. com**. Pakistan Household Income per Capita (2005-2019), Accessed on 19 Aug 2021. Available from: www.ceicdata.com/en/indicator/pakistan/annual-household-income-per-capita

52 **Salazar CR**, Francois F, Li Y, Corby P, Hays R, Leung C, Bedi S, Segers S, Queiroz E, Sun J, Wang B, Ho H, Craig R, Cruz GD, Blaser MJ, Perez-Perez G, Hayes RB, Dasanayake A, Pei Z, Chen Y. Association between oral health and gastric precancerous lesions. *Carcinogenesis* 2012; **33**: 399-403 [PMID: 22139442 DOI: 10.1093/carcin/bgr284]

53 **Uno Y**. Prevention of gastric cancer by Helicobacter pylori eradication: A review from Japan. *Cancer Med* 2019; **8**: 3992-4000 [PMID: 31119891 DOI: 10.1002/cam4.2277]

54 **Song JH**, Kim YS, Heo NJ, Lim JH, Yang SY, Chung GE, Kim JS. High Salt Intake Is Associated with Atrophic Gastritis with Intestinal Metaplasia. *Cancer Epidemiol Biomarkers Prev* 2017; **26**: 1133-1138 [PMID: 28341758 DOI: 10.1158/1055-9965.EPI-16-1024]

55 **Monteiro C**, Costa AR, Peleteiro B. Sodium intake and Helicobacter pylori infection in the early stages of life. *Porto Biomed J* 2016; **1**: 52-58 [PMID: 32258550 DOI: 10.1016/j.pbj.2016.05.001]

56 **Gaddy JA**, Radin JN, Loh JT, Zhang F, Washington MK, Peek RM Jr, Algood HM, Cover TL. High dietary salt intake exacerbates Helicobacter pylori-induced gastric carcinogenesis. *Infect Immun* 2013; **81**: 2258-2267 [PMID: 23569116 DOI: 10.1128/IAI.01271-12]

57 **Park JY**, Forman D, Waskito LA, Yamaoka Y, Crabtree JE. Epidemiology of Helicobacter pylori and CagA-Positive Infections and Global Variations in Gastric Cancer. *Toxins (Basel)* 2018; **10** [PMID: 29671784 DOI: 10.3390/toxins10040163]

58 **World Health Organization**. Reducing salt intake in populations: Rep WHO Forum, Oct 5-7, 2006, Paris, France. Available from: https://apps.who.int/iris/handle/10665/43653

59 **Shi X**, Zhang J, Mo L, Shi J, Qin M, Huang X. Efficacy and safety of probiotics in eradicating Helicobacter pylori: A network meta-analysis. *Medicine (Baltimore)* 2019; **98**: e15180 [PMID: 30985706 DOI: 10.1097/MD.0000000000015180]

60 **Gunathilake MN**, Lee J, Jang A, Choi IJ, Kim YI, Kim J. Physical Activity and Gastric Cancer Risk in Patients with and without Helicobacter pylori Infection in A Korean Population: A Hospital-Based Case-Control Study. *Cancers (Basel)* 2018; **10** [PMID: 30279385 DOI: 10.3390/cancers10100369]

61 **Ayala G**, Escobedo-Hinojosa WI, de la Cruz-Herrera CF, Romero I. Exploring alternative treatments for Helicobacter pylori infection. *World J Gastroenterol* 2014; **20**: 1450-1469 [PMID: 24587621 DOI: 10.3748/wjg.v20.i6.1450]

62 **Monno R**, De Laurentiis V, Trerotoli P, Roselli AM, Ierardi E, Portincasa P. Helicobacter pylori infection: association with dietary habits and socioeconomic conditions. *Clin Res Hepatol Gastroenterol* 2019; **43**: 603-607 [PMID: 30905666 DOI: 10.1016/j.clinre.2018.10.002]

63 **Choi Y**, Choi HS, Jeon WK, Kim BI, Park DI, Cho YK, Kim HJ, Park JH, Sohn CI. Optimal number of endoscopic biopsies in diagnosis of advanced gastric and colorectal cancer. *J Korean Med Sci* 2012; **27**: 36-39 [PMID: 22219611 DOI: 10.3346/jkms.2012.27.1.36]

64 **Centers for Disease Control (CDC)**. Smoking and cancer. *MMWR Morb Mortal Wkly Rep* 1982; **31**: 77-80 [PMID: 6801462]

Published by **Baishideng Publishing Group Inc**
7041 Koll Center Parkway, Suite 160, Pleasanton, CA 94566, USA
Telephone: +1-925-3991568
E-mail: bpgoffice@wjgnet.com
Help Desk: https://www.f6publishing.com/helpdesk
https://www.wjgnet.com