Name of Journal: World Journal of Gastrointestinal Oncology
Manuscript NO: 76367
Manuscript Type: EVIDENCE-BASED MEDICINE

Prediction of gastric cancer risk by a polygenic risk score of *Helicobacter pylori*

Wang XY *et al.* PRS of *H. pylori*
Abstract

BACKGROUND
Genetic variants of *Helicobacter pylori* (*H. pylori*) are involved in gastric cancer occurrence. Single nucleotide polymorphisms (SNPs) of *H. pylori* that are associated with gastric cancer have been reported. The combined effect of *H. pylori* SNPs on the risk of gastric cancer remains unclear.

AIM
To assess the performance of a polygenic risk score (PRS) based on *H. pylori* SNPs in predicting the risk of gastric cancer.

METHODS
A total of 15 gastric cancer-associated *H. pylori* SNPs were selected. The associations between these SNPs and gastric cancer were further validated in 1022 global strains with publicly available genome sequences. The PRS model was established based on the validated SNPs. The performance of the PRS for predicting the risk of gastric cancer was assessed in global strains using quintiles and random forest (RF) methods. The variation in the performance of the PRS among different populations of *H. pylori* was further examined.

RESULTS
Analyses of the association between selected SNPs and gastric cancer in the global dataset revealed that the risk allele frequencies of six SNPs were significantly higher in gastric cancer cases than non-gastric cancer cases. The PRS model constructed subsequently with these validated SNPs produced significantly higher scores in gastric cancer. The odds ratio (OR) value for gastric cancer gradually increased from the first to the fifth quintile of PRS, with the fifth quintile having an OR value as high as 9.76 (95% confidence interval: 5.84-16.29). The results of RF analyses indicated that the area under the curve (AUC) value for classifying gastric cancer and non-gastric cancer was 0.75,
suggesting that the PRS based on *H. pylori* SNPs was capable of predicting the risk of gastric cancer. Assessing the performance of the PRS among different *H. pylori* populations demonstrated that it had good predictive power for cancer risk for hpEurope strains, with an AUC value of 0.78.

**CONCLUSION**
The PRS model based on *H. pylori* SNPs had a good performance for assessment of gastric cancer risk. It would be useful in the prediction of final consequences of the *H. pylori* infection and beneficial for the management of the infection in clinical settings.

**Key Words:** Polygenic risk scores; *Helicobacter pylori*; Gastric cancer; Single nucleotide polymorphism


**Core Tip:** Prediction of cancer risk is of importance in the clinical management of populations with a high risk of gastric cancer. This study constructed a polygenic risk score (PRS) model based on *Helicobacter pylori* (*H. pylori*) single nucleotide polymorphisms (SNPs) to predict the risk of gastric cancer. Associations between previously reported *H. pylori* SNPs and gastric cancer were validated in global strains. A PRS model constructed with validated SNPs had a high predictive power for gastric cancer at a global level and for individuals infected with hpEurope strains. It has potential for clinical use in the management of the *H. pylori* infection.

**INTRODUCTION**
*Helicobacter pylori* (*H. pylori*) infection affects more than half of the world’s population[1,2]. The outcomes of *H. pylori* infection vary among individuals. The
consequences of most infections are benign. However, a minority of infected individuals may eventually develop gastric cancer\textsuperscript{[3,4]}. Predicting the outcomes of \textit{H. pylori} infection is a major concern in the management of the infection. Substantial genetic variation has been found in the pathogen. Mutations cause increased virulence in certain strains, enhancing their carcinogenic potential\textsuperscript{[5,6]}. It has been demonstrated that typing \textit{H. pylori} strains based on the genetic variations of virulent genes has the potential to predict the risk of gastric cancer\textsuperscript{[7,8]}.

Two studies have been recently conducted to investigate the association between \textit{H. pylori} genomic variations and gastric cancer within the hpEurope and hpEastAsia populations, respectively\textsuperscript{[9,10]}. The first study contained 173 hpEurope strains and found 11 cancer risk-associated variants, including gene loss variants and single nucleotide polymorphisms (SNPs). Risk scores calculated based on the status of the \textit{cag}11, \textit{cag}12 and \textit{cag}20 genes were increased during the progression from inflammation to gastric cancer. The other study identified 11 SNPs and three DNA motifs associated with gastric cancer through examination of 240 hpEastAsia strains. It is unclear whether the association between these variations and gastric cancer exists for all \textit{H. pylori} strains. However, the findings from these studies suggest that SNPs from the \textit{H. pylori} genome have the potential to predict the risk of gastric cancer.

To explore the combined effect of multiple SNPs on disease susceptibility, the polygenic risk score (PRS) model has been developed\textsuperscript{[11]}. A PRS is calculated as a sum of the effects of multiple SNPs on disease. PRS models composed of SNPs from the human genome have been successfully used to predict the risk of cancers such as gastric cancer, colorectal cancer, and breast cancer\textsuperscript{[12-15]}. Few studies, however, have been conducted to explore the capacities of PRS model constructed with SNPs from bacterial genomes in predicting the risk of cancer. Our study aimed to construct a PRS model based on validated risk alleles of \textit{H. pylori} to predict the risk of gastric cancer.

**MATERIALS AND METHODS**

\textit{Strains and SNP selection}
A total of 2022 *H. pylori* genome sequences deposited in GenBank at the National Center for Biotechnology Information by December 8, 2021 (https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/169/), and the figshare website (https://figshare.com/s/2174da1fa20ae71c71e0)[10] were downloaded for further analyses. Of them, 1187 *H. pylori* strains had relevant clinical information of patients. Subsequently, duplicate strains and strains isolated from peptic ulcer disease, mucosa-associated lymphoma or stromal tumors were excluded from further analyses. This led to a final dataset of 1022 global strains included in the study. They were divided into gastric cancer (n = 253) and non-gastric cancer (n = 769) groups. Patients in the latter group were diagnosed with functional dyspepsia (n = 46), or chronic gastritis with or without intestinal metaplasia (n = 143 and n = 580, respectively). A total of 15 *H. pylori* SNPs or genetic variants from the two previous genome-wide association studies (GWASs) were selected for further analyses (Table 1)[8,10]. The selection criteria were as follows: (1) The length of the variants was no longer than five contiguous nucleotides; and (2) The SNP selected was located in a protein-coding region.

**Construction of the neighbour-joining tree**

Based on the 1022 *H. pylori* genomes, the SNPs in the core genome (present in > 99% isolates) were identified by aligning the assembled genomes against the reference genome (26695-1MET, accession number: CP010436.1) using MUMmer as previously described[16,17]. A neighbour-joining tree was then constructed based on the sequences of concatenated SNPs using TreeBeST software (http://treesoft.sourceforge.net/trebest.shtml) with default parameters.

**Statistical analyses**

The chi-square test was used to test the difference in the prevalence of risk alleles in strains isolated from gastric cancer and non-gastric cancer. Student’s *t* test was used to compare the PRS values between the gastric cancer and non-cancer groups. These tests were performed using SPSS 18.0 software. Odds ratios (ORs) and 95% confidence
intervals (CIs) of the selected SNPs were calculated using logistic regression analysis in R (version 3.6.3).

A PRS was created for each strain using the following equation: \( \text{PRS} = \beta_1 + \beta_2 + \ldots + \beta_k + \ldots + \beta_n \). Briefly, in this equation, \( \beta_k \) is the value obtained from the regression analysis of the risk allele and disease, and \( n \) is the total number of SNPs included in the PRS\[18\]. Logistic regression analysis was performed to evaluate the association between PRS and gastric cancer risk and by quintiles of the PRS risk distribution, standardized by the controls, and using the 3rd quintile, 40%-60%, as the reference\[18\].

**Random forest algorithm**

A random forest (RF) model was built using the AUC-RF algorithm\[19\]. The input variables were the scores of each of the validated SNPs. A 20-times repeated 10-fold cross-validation of the RF model was performed. The performance of the RF model was demonstrated by receiver operating characteristic curve analysis\[20\].

**RESULTS**

**Validation of SNPs in the global dataset**

Previous studies have identified two sets of *H. pylori* SNPs that are associated with gastric cancer\[9,10\]. The association between these SNPs and gastric cancer has been verified only in strains from the hpEurope or hpEastAsia populations, respectively. We selected 15 SNPs to validate the association between selected SNPs and gastric cancer in global strains (Table 1). The risk alleles were defined as those with a higher prevalence in strains from gastric cancer. Statistical analyses revealed that the risk alleles of six SNPs showed a significant increase in prevalence in the gastric cancer group compared with the non-gastric cancer group. These SNPs, validated in the global dataset, were used for subsequent analyses.

**Establishment of the six-SNP PRS model**
To construct a PRS model for predicting the risk of gastric cancer, the logOR values of each validated SNP were calculated (Table 1). A PRS model was subsequently constructed with the sum of the logOR values of six validated SNPs. The mean PRS value was 8.64 ± 1.71 and 6.99 ± 1.27 in the gastric cancer and non-gastric cancer groups, respectively. The PRS value in the gastric cancer group was significantly higher ($P = 5.6\text{E-}36$).

*Evaluation of the association between PRS and gastric cancer risk*

To evaluate the performance of the 6-SNP PRS model for predicting the risk of gastric cancer, the PRS values for each of the selected 1022 strains were grouped according to the quintile method. With the third quintile as the reference, the estimated OR value gradually increased from the first quintile (< 20%) to the fifth quintile (> 80%) (Figure 1, Table 2). The fifth quintile had an OR value as high as 9.76 (95% CI: 5.84-16.29).

To further confirm the combined effect of the validated SNPs for prediction of gastric cancer risk, an RF model was constructed with logOR values from each SNP as input. The classification potentials of the combined logOR values of validated SNPs were then analysed. The importance of each SNP is shown in Figure 2. The AUC value was 0.75 (DeLong 95% CI: 0.71-0.78), suggesting a good classifying capacity of the combined SNPs.

*Performance of risk prediction by PRS for different H. pylori populations*

Considering the remarkable genomic variations among strains from different *H. pylori* populations, the performance of PRS for predicting the risk of gastric cancer was subsequently assessed in different *H. pylori* populations. The results of the phylogenetic analyses divided the 1022 global strains into five groups, namely, the hpEastAsia, hpAsia2, hpEurope, America-related and Africa-related populations (Figure 3). Due to the small number of gastric cancer cases (2 cases in hpAsia2 and no cases in Africa-related populations), hpAsia2 and Africa-related populations were excluded from subsequent analyses. In analysing the performance of the established PRS model in
different populations, the PRS value was higher in the gastric cancer group for all populations. Statistical analyses revealed a significant difference in PRS between the gastric cancer and non-gastric cancer groups in the hpEastAsia, hpEurope and America-related populations (Figure 4).

To further verify the combined effects of these SNPs for prediction of gastric cancer risk for different H. pylori populations, a RF classification model was built. The results of RF model analyses demonstrated that the AUC value was highest (0.78, DeLong 95%CI: 0.70-0.85) in the hpEurope population, suggesting a good ability of the combined SNPs to predict the risk of gastric cancer (Figure 5). However, the performance of the combined SNPs for risk prediction in other H. pylori populations was poor (Figure 5).

**DISCUSSION**

In this study, we constructed a PRS model based on validated H. pylori SNPs to predict the risk of gastric cancer. To our knowledge, our study is the first to evaluate a PRS model for cancer risk prediction constructed with genomic variants of H. pylori. H. pylori shows substantial genetic variations, resulting in remarkable interstrain differences in carcinogenetic potential\[5,21\]. The presence/absence or large sequence variation of virulence genes and H. pylori SNPs have been shown to promote gastric carcinogenesis. Few studies have been conducted to assess the predictive power of these cancer-related genetic variations for gastric cancer\[9,10\]. Moreover, the combined effect of multiple variations on the predictive power for cancer risk has not been explored. Findings from this study demonstrate that a PRS model combining six H. pylori SNPs had a moderate capacity for prediction of gastric cancer risk. This is similar to the findings in studies on PRS model constructed with cancer-associated SNPs from the human genome\[14,15\].

To assess the combined effects of SNPs on gastric cancer risk prediction, we first selected 15 cancer-associated H. pylori SNPs from two previous GWAS studies. Their association has been validated in strains from specific geographical regions but not in a global strain collection. Our results demonstrated that only six of the SNPs showed a
close association with gastric cancer in the global dataset. The SNPs at 88029, 241625, 803467 and 854415 in the reference strain 26695 caused nonsynonymous changes in the corresponding amino acid sequence, whereas the SNPs at 140797 and 1117402 in the reference strain 26695 produced synonymous variations. The \textit{hpaA} gene, harbouring the SNP at 854415, encodes an adhesion gene of \textit{H. pylori}\cite{32}. This gene is essential for colonization and is associated with the occurrence of gastric cancer\cite{23, 25}. The SNP at 88029 was located on the \textit{tlpC} gene. \textit{TlpC} encodes a chemoreceptor that affects the chemotaxis of strains in the mouse gastric environment. It is associated with the induction of mucosal inflammation of the stomach\cite{26, 27}. The SNP at 241625 was located in \textit{dsbG/K}, which has protein disulfide isomerase activity. \textit{DsbG/K} interacts with a virulence-related factor \textit{in vitro}\cite{28, 29}. \textit{In vitro} studies have shown that a lack of \textit{dsbG/K} may cause the loss of T4SS function and inhibit \textit{VacA} secretion, which are considered the main pathogenic factors in \textit{H. pylori}\cite{30}.

In this study, we constructed a PRS model with six SNPs validated in a global dataset. Assessments of the performance of the PRS model demonstrated that the PRS value was significantly higher in the gastric cancer group than in the non-gastric cancer group. A significant increase in the risk of gastric cancer was found across the quintiles of the PRS. These findings demonstrate that the six-SNPs PRS model is capable of predicting the risk of gastric cancer. In support of this finding, RF analyses demonstrated that the combination of the six SNPs has a high predictive power for gastric cancer, with an AUC value of 0.75. In a recent report, a PRS model constructed with SNPs from the human genome showed unsatisfactory power in classifying gastric cancer from healthy controls, with an AUC value of 0.56\cite{31}. It has been shown that a PRS model derived from 112 SNPs in the human genome and lifestyle factors possesses good predictive capacity for gastric cancer risk\cite{32}. For individuals infected with \textit{H. pylori}, assessment of their gastric cancer risk is of great concern in the clinical settings. Previous reports have demonstrated that certain genetic variants are associated with increased gastric cancer risk\cite{9, 10}. Our study, for the first time, demonstrated the combined effect of \textit{H. pylori} genomic variations in the assessment of cancer risk. The
PRS model derived from *H. pylori* SNPs would have a high capacity in predicting gastric cancer risk for patients infected with the pathogen. This will benefit the clinical management of the prognosis of the *H. pylori* infection. It is well known that age, gender and lifestyle factors, including alcohol consuming, smoking, diet habits and economic status, are closely associated with gastric cancer[33-35]. In the future, a PRS model constructed with *H. pylori* SNPs and those gastric cancer associated risk factors in this study would have substantially increased power in predicting the risk of gastric cancer. The *H. pylori* genome shows great variations between strains[36,37]. Genetic information differs greatly among *H. pylori* populations, and their carcinogenic potential is also different[8,21]. We thus evaluated the performance of the PRS model across *H. pylori* populations. Our results demonstrated a good predictive power of PRS for hpEurope strains.

A limitation of this study is that the performance of the PRS model was not assessed in hpAsia2 and Africa-related *H. pylori* populations because the number of strains with clinical information available was insufficient. Moreover, we could not consider age, gender, nutrition and other risk factors in the construction of the PRS model, as information on all of these risk factors was not consistently available across databases. A comprehensive risk model enclosing other risk factors of gastric cancer is indicated in future studies. Further *in vitro* and *in vivo* exploration of the roles of the combination of *H. pylori* SNPs identified in this study in gastric cancer would be much helpful in supporting our findings.

**CONCLUSION**

In summary, we constructed a PRS model based on *H. pylori* SNPs, which showed great potential in the prediction of gastric cancer risk globally, especially for individuals infected with hpEurope strains. Findings from this study demonstrated that the PRS model constructed from bacteria genomic variations, in addition to the PRS model established with human SNPs, can be of great value for disease risk prediction. In clinical practice, it is usually difficult to assess gastric cancer risk in patients infected
with *H. pylori*. The model constructed in this study would be beneficial for solving this issue.

**ARTICLE HIGHLIGHTS**

*Research background*

Multiple single nucleotide polymorphisms (SNPs) of *Helicobacter pylori* (*H. pylori*) associated with gastric cancer have been identified through bacterial genome-wide association studies. Polygenic risk score (PRS) calculated as a sum of effect of SNPs provides a tool for assessing genetic impact on diseases.

*Research motivation*

Predicting risk of gastric cancer is a major concern in the management of the *H. pylori* infection.

*Research objectives*

This study constructed a PRS model based on *H. pylori* SNPs to predict the risk of gastric cancer.

*Research methods*

Associations between previously reported *H. pylori* SNPs and gastric cancer were validated in global strains. The PRS model based on the validated SNPs was evaluated by quintiles and RF methods.

*Research results*

A PRS model was constructed with six validated SNPs. Quintiles and RF methods demonstrated the combination of six SNPs has a high predictive power for gastric cancer.

*Research conclusions*
PRS model constructed from bacterial genomic variations can be of great value for gastric cancer risk prediction.

Research perspectives
Comprehensive risk models including personal and genomic information need to be established in future studies.
<table>
<thead>
<tr>
<th>#</th>
<th>Source</th>
<th>Title</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>patents.google.com</td>
<td>Hiroki Kiyokawa, Masatoshi Abe, Takahiro Matsui, Masako Kurashige et al. &quot;Deep learning analysis of histologic images from intestinal specimen reveal &quot;adipocyte shrinkage&quot; and mast cell infiltration to predict post-operative Crohn's disease.&quot;, The American Journal of Pathology, 2022</td>
<td>&lt; 1%</td>
</tr>
</tbody>
</table>