Gastric Cancer (2018) 21:196–203
https://doi.org/10.1007/s10120-017-0748-z

ORIGINAL ARTICLE

Toward early detection of *Helicobacter pylori*-associated gastric cancer

Rachel Walker[1] · Jan Poleszczuk[2] · Jaime Mejia[3] · Jae K. Lee[4] · Jose M. Pimiento[5] · Mokenge Malafa[5] · Anna R. Giuliano[8] · Heiko Enderling[1,9] · Domenico Coppola[6,7]

Received: 19 March 2017 / Accepted: 8 July 2017 / Published online: 19 July 2017

Abstract

Background Gastric cancer is typically diagnosed at a late stage, leading to poor prognoses. *Helicobacter pylori* is responsible for 70% of gastric cancers globally, and patients with this bacterial infection often present with early stages of the carcinogenic pathway such as inflammation or gastritis. Although many patients continue to progress to advanced-stage disease after antibacterial treatment, there are no follow-up screening protocols for patients with a history of *H. pylori.*

Methods Several biomarkers (Lgr5, CD133, CD44) become upregulated during gastric carcinogenesis. A logistic regression model is developed using clinical data from 59 patients at different stages of the carcinogenic pathway to identify the likelihood of being at an advanced stage of disease for all combinations of age, sex, and marker positivity. Using these likelihood distributions and the observed rate of marker positivity increase, time to high likelihood (probability >0.8) of advanced disease for individual patients is predicted.

Results A strong correlation between marker positivity and disease stage was found for all three markers. Disease stage was accurately classified by the respective regression models for more than 86% of retrospective patients. Highly patient-specific predictions of time to onset of dysplasia were made, allowing the classification of 17 patients initially diagnosed with intestinal metaplasia into high-, intermediate-, or low-risk categories.

Conclusions We present an approach designed to integrate pathology, mathematics, and statistics for detection of the earliest precancerous, treatable lesion. Given the simplicity and robustness of the framework, such technique has the potential to guide personalized screening schedules to minimize the risk of undetected malignant transformation.

Keywords *Helicobacter pylori* · Stomach neoplasms · Early detection of cancer

Electronic supplementary material The online version of this article (doi:10.1007/s10120-017-0748-z) contains supplementary material, which is available to authorized users.

✉ Domenico Coppola
Domenico.Coppola@moffitt.org

[1] Department of Integrated Mathematical Oncology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL 33612, USA

[2] Nalecz Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences, Warsaw, Poland

[3] Instituto de Patología Mejía Jiménez, Cali, Colombia

[4] Department of Biostatistics and Bioinformatics, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL 33612, USA

[5] Department of Gastro Intestinal Oncology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL 33612, USA

[6] Department of Anatomic Pathology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL 33612, USA

[7] Department of Tumor Biology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL 33612, USA

[8] Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL 33612, USA

[9] Department of Radiation Oncology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL 33612, USA

Introduction

At present, as many as 80% of gastric cancer patients reach stage IV before clinical diagnosis [1]. To reduce the particularly high mortality rate of these cancers, early intervention is essential. Although it is known that several biomarkers become upregulated during gastric carcinogenesis [2–4], a concerted effort is still needed to thoroughly evaluate their clinical applicability. Here, we show how these biomarkers may facilitate the development of a screening methodology for the early detection of preneoplastic lesions.

Helicobacter pylori is responsible for up to 70% of gastric cancers worldwide, initiating a carcinogenic cascade from chronic active gastritis to intestinal metaplasia, dysplasia, and ultimately carcinoma: the Correa pathway [5]. Symptoms of the early stages of this pathway are common because of the extensive inflammation and tissue damage induced by bacterial colonization, and often lead patients to the clinic at the chronic gastritis or metaplasia stage. However, despite a continued risk of progression through the carcinogenic pathway (because of damage that is not reversed following bacterial eradication [6–8]), no follow-up screening is routinely provided. Patients who continue to progress predominantly return to the clinic when the disease is already invasive or metastatic, at which point curative treatment is a near impossibility. If patients with a history of *H. pylori* infection underwent periodic follow-up, they could benefit from the early detection of progression to low-grade dysplasia. This program would allow endoscopic monitoring and early surgical intervention, with the promise of significantly improved outcomes. As such, the aim of this study is to demonstrate how the quantitative tools of mathematics and statistics may complement known biomarkers of disease progression and contribute to the development of effective screening protocols.

Lgr5, CD44, and CD133 are upregulated in gastric cancer tissue [3, 4], and their expression levels have several implications for metastasis, therapy resistance, and overall prognosis [9–11]. Expression of these markers also increases incrementally between each respective stage of the Correa pathway [2]. This change suggests these markers could be used as indicators of patient-specific disease progression and could be incorporated into predictive tools designed to optimize follow-up screening schedules for patients with a history of *H. pylori* infection.

Here, we investigated if statistical models could determine the likelihood of a patient being either early in the Correa pathway (gastritis or metaplasia) or late in the pathway (low- or high-grade dysplasia or carcinoma) based on patient age, sex, and biomarker-positive cell fraction (obtained from immunohistochemical staining of tissue samples). These models are calibrated using an initial cohort of patients at different stages of disease, and model ability to accurately classify tissue samples is demonstrated by comparing model predictions to disease stages determined by pathology. If the rate of increase of the marker-positive cell population during disease progression can be derived from existing clinical data, such a model can be used to identify the time at which the risk of progressing to low-grade dysplasia will be above a certain threshold for that patient. This time can be used to classify an individual patient's current risk status dependent on their patient-specific input parameters (marker positivity, age, sex), and recommend follow-up screening at an optimal stage in the pathway: late enough to avoid frequent and costly overscreening, but early enough for treatment to have a high likelihood of success.

Utilizing candidate biomarkers of disease progression to improve actual clinical outcomes is a necessarily multidisciplinary challenge. The purpose of the present work is to demonstrate the potential for a quantitative framework to contribute to bridging this gap, toward early detection and intervention in cancers with typically late diagnoses.

Materials and methods

Patients and sample

Retrospective gastric biopsy samples were collected from 59 *H. pylori*-positive patients from the Instituto de Patologia Mejia Jimenez in Cali, Colombia during 2014 and shipped as formalin-fixed paraffin-embedded (FFPE) tissue blocks to Moffitt Cancer Center, Tampa, FL (USA). *H. pylori* infection prevalence is approximately 70% in Colombia, reaching even higher levels in populations residing in the south of the country and in mountainous regions [12]. This locale provides an optimal site for the analysis of *H. pylori*-associated disease. The prevalent bacterial strain in this region is CagA+/VacA+, the strain typically associated with carcinogenesis; because of the history of the bacterial infection, gastric cancer (GC) patients primarily presented with intestinal-type gastric cancers of the antrum and corpus. Disease stage was assessed by hematoxylin and eosin (H&E) staining, and *H. pylori* status was evaluated by immunohistochemical staining of endoscopy samples from each patient. Patients were selected from four different stages of the Correa pathway based on pathological analysis of histological lesions: normal gastric mucosa (NM), complete intestinal metaplasia (IM), low-grade dysplasia (DS), and adenocarcinoma (GC). Baseline patient characteristics are summarized in Table 1.

Immunohistochemistry

All samples were stained for three putative carcinogenesis biomarkers (Lgr5, CD44, CD133). A 4-μm section of all selected blocks was stained using a Ventana Discovery XT automated system (Ventana Medical Systems, Tucson, AZ, USA) as per manufacturer's protocol with proprietary reagents. The antibodies used were the rabbit anti-human LGR5 primary antibody (ab75850; Abcam, Cambridge, MA, USA; 1:100 dilution), rabbit anti-human CD44 primary antibody (#HPA005785; Sigma Aldrich, St. Louis, MO, USA; 1:1000 dilution), and mouse anti-human CD133 monoclonal antibody (MAB4399; Millipore, Billerica, MA, USA; 1:100 dilution). Stained slides were read by two independent pathologists. Marker positivity was quantified by fraction of epithelial cells staining. Inflammatory cell staining was not included in marker positivity scores.

Regression modeling

Logistic regression models based on several predictors (age, sex, and biomarker-positive cell fraction) were developed for estimating the probability of an individual being at either early stage (gastritis or metaplasia, stage <DS) or late stage (low-grade dysplasia or carcinoma, stage ≥DS). Note that high-grade dysplasia was observed in the surrounding tissue of several samples of gastric carcinoma; however, the most advanced histological lesion visible on the sample was used to classify stage for model development. The transition from metaplasia to low-grade dysplasia is a clinically relevant and clearly histologically defined timepoint, and if identified early could allow further clinical action in the form of endoscopic monitoring or surgical intervention.

Model fitting was conducted using the binomial family of the inbuilt generalized linear model (GLM) fitting function of statistical software R; a series of regression coefficients was generated from which it was possible to evaluate the contribution of each of the respective predictors to outcome. The Wilcoxon rank-sum test with a normal approximation was used to compare marker positivity between males and females and age between males and females. The independence of continuous variables age and marker-positive fraction was tested using Pearson's product-moment correlation test. The significance level was set at p value = 0.05. Given the relatively small sample size, jackknife resampling was conducted to evaluate the generalizability of the observed regression coefficients and subsequent model predictions. All regression modeling and statistical analyses were performed using R software.

Classification performance

Given patient-specific input information (age, sex, marker positivity), the model generated the likelihood that the patient is at either stage <DS or stage ≥DS. For our initial cohort of 59 patients, the ability of the model to correctly classify patients into the category matching their pathology report was evaluated, and model under- and over-predictions were scored.

Follow-up screening times

Based on this regression modeling, we calculate the probability of a patient being at an advanced stage of disease (stage ≥DS) based on their marker-positive fraction and age for both females and males. From the derived probability distributions and initial patient-specific characteristics, a mathematical model of marker-positive fraction increase is used to predict the time until the risk of progressing to low-grade dysplasia is above a specific threshold (here set to 80%). For demonstrative purposes, the growth rates of the three markers were approximated based on simple first- and second-order curve fitting to the average marker-positivity data. Time for progression from clinically symptomatic gastritis to gastric cancer of 2 years was selected based on the average progression time of 12 patients in an independent cohort from the same institution.

We derived model-predicted times until the likelihood of low-grade dysplasia reaches 80% for all realistic

Table 1 Baseline characteristics of the studied cohort

	Normal ($n = 10$)	Intestinal metaplasia ($n = 17$)	Dysplasia ($n = 10$)	Gastric cancer ($n = 21$)
Age (years)	43.3 ± 11.1	61.9 ± 16.3	51.7 ± 16.1	53.4 ± 14.3
Sex				
Male	1	9	3	11
Female	9	8	7	10
Marker-positive function (%)				
LGR5	1.3 ± 0.5	9.8 ± 8.5	25 ± 12.7	56.1 ± 24
CD44	1.6 ± 0.9	12.4 ± 7.7	24.2 ± 9.6	42.8 ± 16.8
CD133	0.6 ± 0.5	2.7 ± 1.4	6.9 ± 4.1	25.6 ± 14.9

combinations of patient ages and marker-positive fractions at clinical presentation. From these times, patients could be classified as "high risk" (predicted time to progression <100 days), "intermediate risk" (predicted time to progression between 100 days and 1 year), or "low risk" (predicted time to progression >1 year). These classifications were compared for all three markers to evaluate prediction robustness and verify independent biomarker choice.

Results

Clinical data

Immunohistochemical analysis identified a significant stepwise increase in marker-positive fraction between each stage of disease for all three markers (Fig. 1). The intensity of immunoreactivity also increased (from weak to strong) during the progression from normal tissue to metaplasia, dysplasia, and cancer. Immunoreactivity for CD44 and CD133 was localized to the cytoplasm, and the LGR5 stain had membranous localization. For Lgr5 and CD44, higher marker positivity was observed in males (Supplementary Fig. 1). As sex and marker positivity are not independent predictors, all further analysis was conducted independently for male and female cohorts. There was no statistically significant difference in age between males and females (Supplementary Fig. 2), or correlation between age and marker positivity for any of the three markers (Supplementary Fig. 3). Correlations between the three respective markers at each stage of disease and corresponding standard deviations are provided in Supplementary Table 2A, B.

Regression modeling

Regression models were fitted to patient data to evaluate the association of patient-specific characteristics with outcome (stage <DS). Table 2 shows model coefficients and intercepts for each of the three respective markers. In five of six logistic regressions, biomarker positivity demonstrates a strong and statistically significant association with probability of being at stage <DS (Table 2), except for CD133 in males. Negative coefficients imply that greater marker positivity decreases the likelihood of being in an early stage. As increasing marker positivity positively correlates with advancing disease stage, this was to be expected. No statistically significant association with probability of outcome was found with age.

Resampling the cohort by systematically omitting each observation and recalculating the regression coefficients and corresponding probabilities demonstrated that the majority of model outputs remained approximately constant (within 10% of the initial calculation) (Supplementary Fig. 8). However, for as many as three patients in each group (LGR5, CD133, and CD44 for both males and females), omission led to coefficients significantly (>10%) different from the initial calculation. In all cases omission of these "outliers" generated a higher correlation of marker expression with disease stage (Supplementary Table 1); in no case did correlation

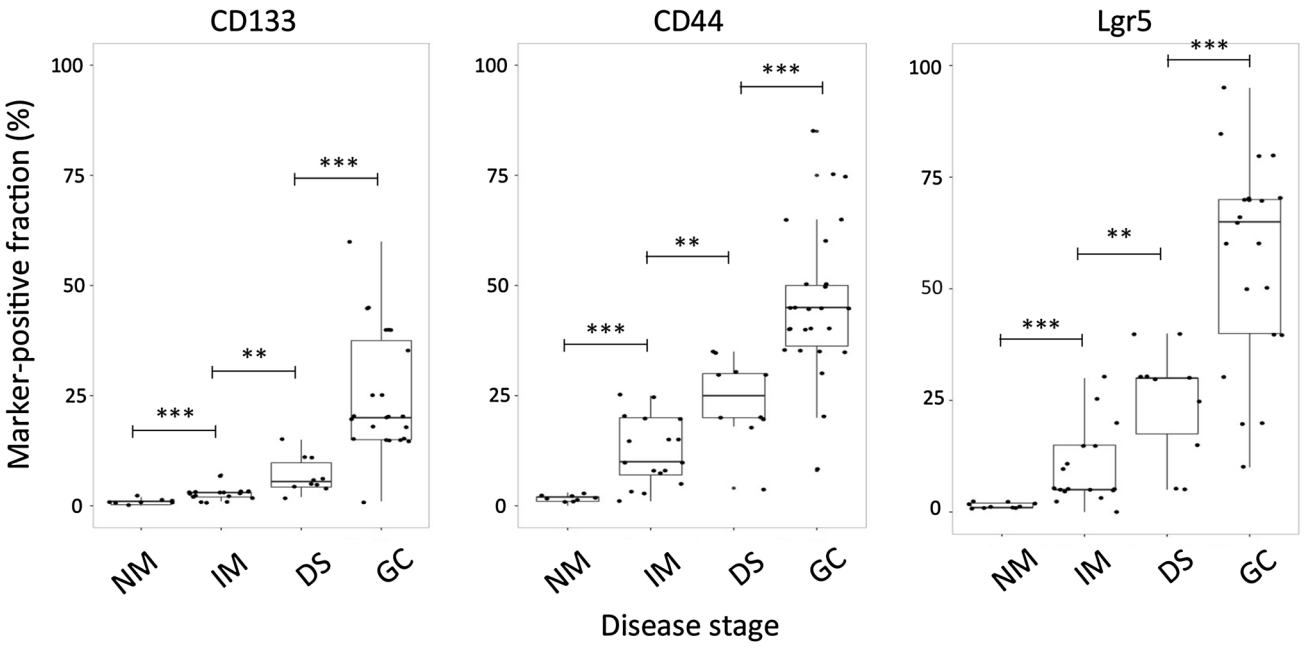

Fig. 1 Average immunohistochemistry marker expression in tissue samples of 59 patients at different gastric cancer disease stages: normal gastric mucosa (*NM*), intestinal metaplasia (*IM*), dysplasia (*DS*), adenocarcinoma (*GC*). **a** Lgr5; **b** CD44; **c** CD133

Table 2 Estimated coefficients from three respective regression models

	Lgr5		CD133		CD44	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Male						
Intercept	1.802	2.48	1.35	3.04	0.72	2.84
Marker	−15.46*	6.88	−76.36	63.93	−19.13*	7.56
Age	0.02	0.04	0.04	0.05	0.05	0.05
Female						
Intercept	0.81	2.55	1.48	2.35	−3.42	3.35
Marker	−24.81*	11.84	−78.91*	33.87	−53.09*	31.87
Age	0.05	0.06	0.04	0.04	0.26	0.17

* p value <0.05; *S.E.* standard error

decrease. This result suggests that the observed positive correlation between marker positivity and disease stage is unlikely to be an artifact of the sample size.

Classification performance

Based on marker positivity, sex, and age, the Lgr5 model correctly classified 49 of 57 cases (86%), with overestimation and underestimation in 4 and 4 cases, respectively. For CD44, the model correctly classified 52 of 59 cases (88%), with overestimation and underestimation in 3 and 4 cases, respectively. The CD133 model correctly classified 53 of 59 cases (91%), with stage overestimation and underestimation in 1 and 4 cases, respectively. Note that each marker was not assessed in an identical number of cases as insufficient tissue was available for some patients.

Follow-up screening times

The predicted probability of a patient being at an advanced stage of disease (low- or high-grade dysplasia or carcinoma, stage ≥DS) based on Lgr5-positive fraction and age for both females and males is shown in Fig. 2, and for CD44 and CD133 in Supplementary Figs. 4 and 5. Based on patient-specific averages, the increase in Lgr5 positivity was linear in males and quadratic in females (Fig. 3). For all combinations of age, sex, and initial marker positivity, iterative increases in the continuous variables were simulated based on these increase rates. The statistical model was used to predict the likelihood of the patient having reached a stage ≥DS. At forecasted likelihoods greater than 80%, a follow-up screen may be suggested. A map of suggested follow-up screening times for all potential combinations of patient input parameters is shown in Fig. 3. Comparable analyses for markers CD44 and CD133 are included in Supplementary Figs. 6 and 7. From these suggestions, patients can be classified into high (predicted time to progression <100 days), intermediate (predicted time to progression between 100 days and 1 year), or low (predicted time to progression >1 year) risk categories. Figure 4 demonstrates that for 8 of the 17 patients from our initial cohort diagnosed with intestinal metaplasia, all three models independently apply the same classification. For a further 8 patients, classifications are in successive categories, in which case the earlier of the two would be used for follow-up recommendation. In only 1 of 17 cases are both low- and high-risk classifications made for the same patient depending on the marker used for evaluation.

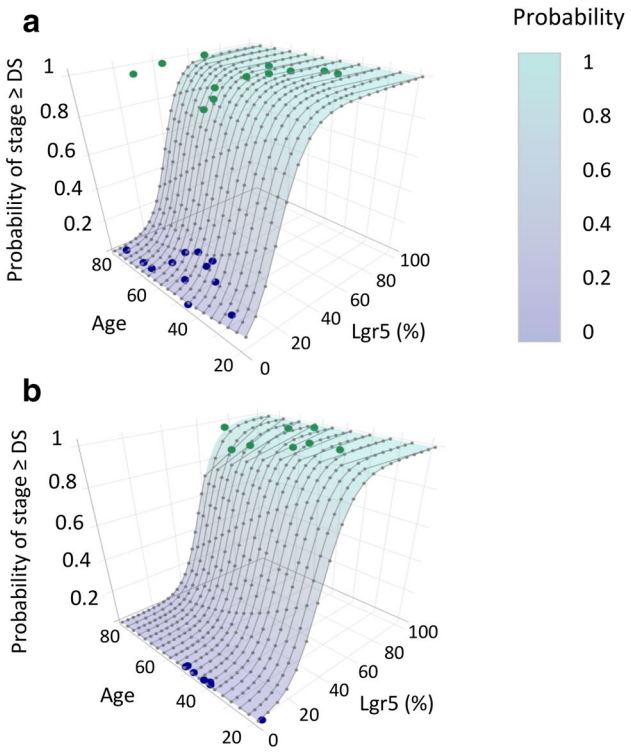

Fig. 2 Model-predicted probability of patient being in advanced stage (stage ≥DS) depending on Lgr5-positive fraction and age for males (**a**) and females (**b**). *Green circles* represent patients at advanced stage (DS or GC); *blue circles* represent patients at early stage [normal gastric mucosa (*NM*) or intestinal metaplasia (*IM*)]

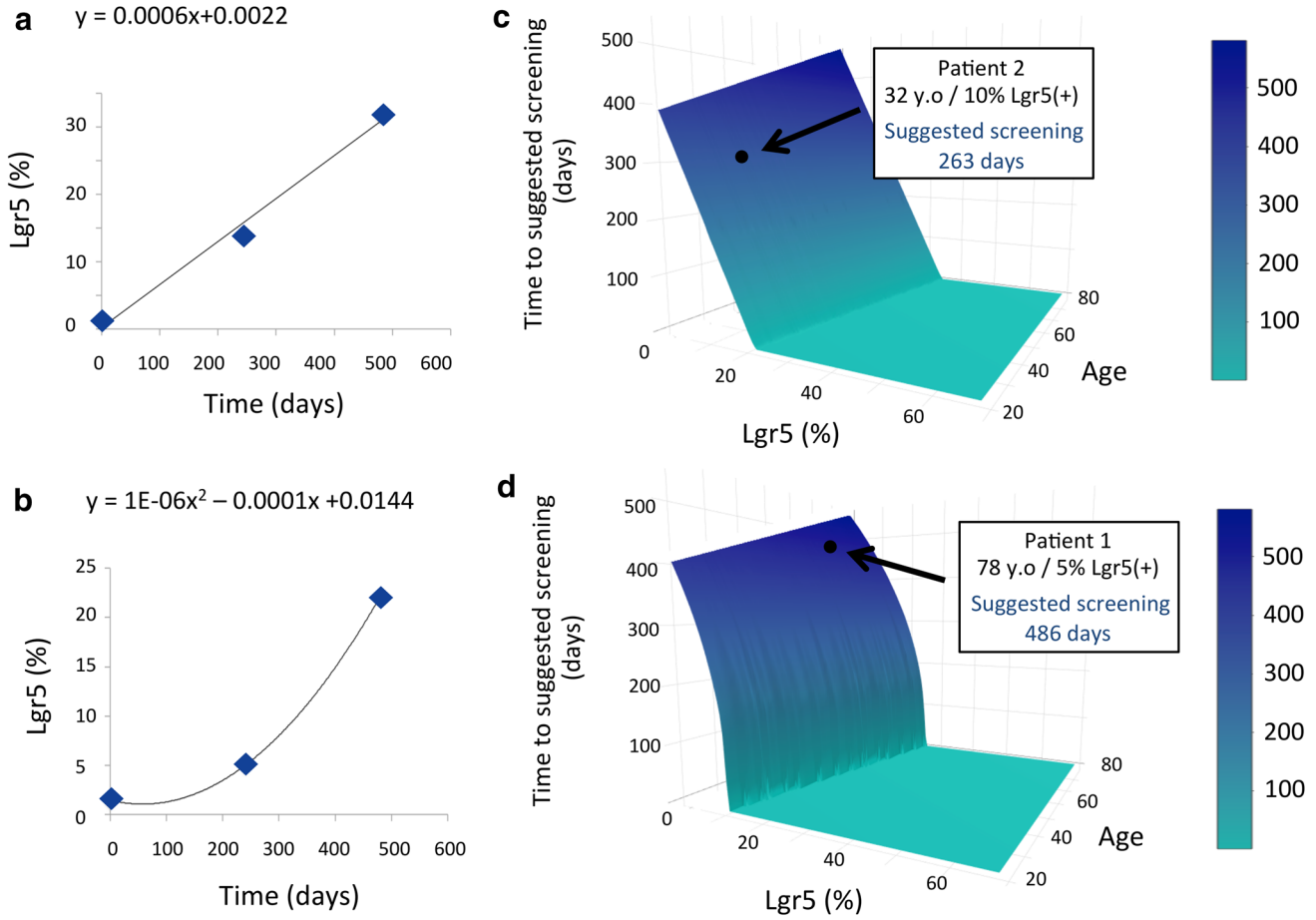

Fig. 3 **a**, **b** Equations approximately governing the increase in positivity of Lgr5 in males (**a**) and females (**b**), respectively, based on patient-specific averages (indicated by *blue diamonds*) obtained in preliminary data (Fig. 1). Model-suggested screening times for males (**c**) and females (**d**), respectively, based on the combined statistical regression tool (Fig. 2) and approximate growth models for the Lgr5-positive cell population (**a**, **b**). Recommended screening time for two examples of male and female patients from the preliminary cohort is highlighted

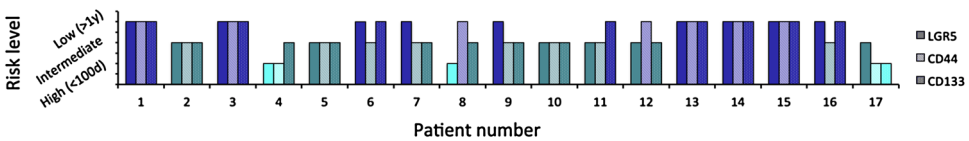

Fig. 4 Model-predicted time to late-stage disease can be used to classify patients as high risk (predicted time <100 days), intermediate risk (predicted time between 100 days and 1 year), and low risk (predicted time >1 year). Classification of each independent model is shown for 17 patients initially diagnosed with intestinal metaplasia, based on immunohistochemical staining of biopsy tissue for three respective markers. From these classifications, clinicians could identify personalized screening protocols based on an individual's current age, sex, and marker positivity

Discussion

Although the high mortality in gastric cancer patients is primarily attributable to late diagnosis, no candidate biomarkers for disease progression are utilized clinically to guide screening for early detection. Several markers have been identified, including Lgr5, CD44, and CD133, that are known to not only be upregulated in gastric cancer but also to increase in a stepwise manner throughout the carcinogenic pathway. This sequential increase introduces the potential for predicting progression through a continuous pathway as opposed to only a binary outcome such as the onset of

neoplasia. However, further effort is needed to find means of utilizing these markers in the clinical setting. The tools of mathematics and statistics have the power to analyze and optimize the use of these markers as predictors of progression, if sufficient data become available with which to calibrate and validate quantitative models.

The statistical tool described here was calibrated with an initial cohort of 59 patients to generate—given a set of patient-specific characteristics—the likelihood the patient is at each respective stage of disease in the Correa pathway. For more than 88% of patient samples, all three of the predefined biomarkers of gastric carcinogenesis when combined with clinical information of patient age and gender were able to accurately predict disease stage of an initial cohort of patients. Only 4% of samples were underscored and 8% of samples were overscored. Without sufficient high-resolution longitudinal data, the growth of the marker-positive cell fraction was fitted to average expression levels across different patients and average disease progression times. As more data on intermediate stages in individual patients become available, more accurate, patient-specific predictive models of marker-positive cell population increase can be derived for higher accuracy predictions.

We demonstrated that increased marker-positive cell fraction during carcinogenesis may be used to identify whether a patient is at high risk of progression to an advanced stage of disease (low-grade dysplasia or later). These classifications can aid clinicians in determining a more cost-effective follow-up screening protocol on a personalized level. Based on these screening recommendations patients can return to the clinic for a follow-up at a time that minimizes the risk of undetected progression to low-grade dysplasia but also does not require excessive and costly overscreening. If low-grade dysplasia or a later stage in the pathway is evident from pathological analysis, the patient can be submitted for closer monitoring or surgical intervention where necessary. If late-stage disease is not yet evident but progression is suggested, either histologically or by a noticeable biomarker increase, a new follow-up screening time may be suggested according to the observed growth rate in marker positivity between the patients' initial presentation and first follow-up. Alternatively, if the patient is experiencing minimal to no persistent inflammation and demonstrates no evidence of progression by increase in marker positivity (marker positivity growth rate approximately zero), it can be assumed that antibacterial triple therapy has been successful and the patient can be removed from the screening protocol at the discretion of the physician.

Although the proposed model provides only approximate screening interval recommendations, under the current paradigm these patients would not be required to undergo any follow-up screening despite the current understanding that eradicating *H. pylori* does not necessarily eradicate gastric cancer risk. The current work presents a proof of concept for an integrated framework to help bridge the gap between candidate biomarkers of disease progression and currently devastating clinical outcomes of late-stage disease. With almost 1 million new cases of gastric cancer each year, a paradigm shift toward early detection and intervention should be a high priority. Importantly, the current approach is not biomarker- or cancer specific; with thorough validation, such frameworks could be applied in the setting of other cancers for which precancerous histological lesions can be monitored and progression-associated biomarkers are established. This approach could have far-reaching implications for the early detection of cancers with typically late diagnoses resulting from currently absent or insufficient, cost-ineffective screening.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Human participants All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1964 and later versions. Informed consent or substitute for it was obtained from all patients for being included in the study. All information linked to the patients is protected and de-identified, and all tissue samples are collected from retrospective cases, thus posing no risks to the patients. Identification was removed from all specimens by the IRB at the Instituto de Patología Mejía Jiménez in Cali, Columbia, both for the initial selection of the samples and for the subsequent data collection.

References

1. Layke JC, Lopez PP. Gastric cancer: diagnosis and treatment options. Am Fam Physician. 2004;69(5):1133–41.
2. Wang T, et al. Sequential expression of putative stem cell markers in gastric carcinogenesis. Br J Cancer. 2011;105:658–65.
3. Nosrati A, Naghshvar F, Khanari S. Cancer Stem Cell Markers CD44, CD133 in primary gastric adenocarcinoma. Int J Mol Cell Med. 2014;3(4):279–86.
4. Zheng ZX, et al. Intestinal stem cell marker LGR5 expression during gastric carcinogenesis. World J Gastroenterol. 2013;19(46):8714–21.
5. Correa P, Haenszel W, Cuello C, Tannenbaum S, Archer M. A model for gastric cancer epidemiology. Lancet. 1975;2(7924):58–60.
6. Asaka M, Kato M, Graham DY. Prevention of gastric cancer by *Helicobacter pylori* eradication. Intern Med. 2010;49:633–6.
7. Graham DY, Shiotani A. The time to eradicate gastric cancer is now. Gut. 2005;54:735–8.
8. Rugge M, Cassaro M, Leo G, Farinati F, Graham DY. *Helicobacter pylori* and gastric cancer: both primary and secondary preventive measures are required. Arch Intern Med. 1999;159:2483–4.
9. Yiming L, et al. CD133 overexpression correlates with clinicopathological features of gastric cancer patients and its impact on survival: a systematic review and meta-analysis. Oncotarget. 2015;6(39):42019.

10. Cho SH, et al. CD44 enhances the epithelial-mesenchymal transition in association with colon cancer invasion. Int J Oncol. 2012;41:211–8.
11. Xi HQ, et al. Leucine-rich repeat-containing G protein-coupled receptor 5 is associated with invasion, metastasis, and could be a potential therapeutic target in human gastric cancer. Br J Cancer. 2014;110:2011–20.
12 Bravo LE, Cortes A, Carrascal E, Jaramillo R, Garcia LS, Bravo PE, Badel A, Bravo PA. *Helicobacter pylori*: paralogia y prevalencia en biopsias gastricas en Colombia. Colombia Med. 2003;34:124–31.