Name of Journal: *World Journal of Gastroenterology*

Manuscript NO: 72837

Manuscript Type: ORIGINAL ARTICLE

Retrospective Cohort Study

Utility of a Deep Learning Model and a Clinical Model for Predicting Bleeding after Endoscopic Submucosal Dissection in Patients with Early Gastric Cancer

Ji Eun Na, Yeong Chan Lee, Tae Jun Kim, Hyuk Lee, Hong-Hee Won, Yang Won Min, Byung-Hoon Min, Jun Haeng Lee, Poong-Lyul Rhee, Jae J. Kim

Abstract

BACKGROUND

Bleeding is one of the major complications after endoscopic submucosal dissection (ESD) in early gastric cancer (EGC) patients. There are limited studies on estimating the bleeding risk after ESD using an artificial intelligence system.

AIM

To derivate and verify the performance of the deep learning model and the clinical model for predicting bleeding risk after ESD in EGC patients.

METHODS

Patients with EGC who underwent ESD between January 2010 and June 2020 at the Samsung Medical Center were enrolled, and post-ESD bleeding (PEB) was investigated retrospectively. We split the entire cohort into a development set (80%) and a validation set (20%). The deep learning and clinical model were built on the development set and tested in the validation set. The performance of the deep learning model and the clinical model were compared using the area under the curve (AUC) and the stratification of bleeding risk after ESD.

RESULTS

A total of 5,629 patients were included, and post-ESD bleeding (PEB) occurred in 325 patients. The AUC for predicting PEB was 0.71 (95%CI, 0.63-0.78) in the deep learning model and 0.70 (95%CI, 0.62-0.77) in the clinical model, without significant difference (*P* = 0.730). The patients expected to the low- (<5%), intermediate- (≥5%, <9%), high-risk (≥9%) categories were observed with actual bleeding rate of 2.2%, 3.9%, and 11.6% in the deep learning model; 4.0%, 8.8%, and 18.2% in the clinical model.

CONCLUSION

A deep learning model can predict and stratify the bleeding risk after ESD in patients with EGC.

Key Words: Clinical model; Deep learning model; Post-ESD bleeding; Stratification of bleeding risk

Na JE, Lee YC, Kim TJ, Lee H, Won HH, Min YW, Min BH, Lee JH, Rhee PL, Kim JJ. Utility of a Deep Learning Model and a Clinical Model for Predicting Bleeding after Endoscopic Submucosal Dissection in Patients with Early Gastric Cancer. *World J Gastroenterol* 2022; In press

Core Tip: Bleeding is one of the major complications after ESD in EGC patients and requires hospital-based intervention. We established a deep learning model to stratify the bleeding risk after ESD and demonstrated its performance compared with a clinical model. The deep learning model showed acceptable AUC and could stratify the PEB risk as low-, intermediate-, and high-risk categories, which correlated with actual bleeding rate comparatively. A deep learning model would be valuable in assessing the bleeding risk after ESD in EGC patients.

INTRODUCTION

In South Korea, gastric cancer has a high incidence, the second most common malignancy, and the fourth most common cause of cancer-related mortality[1]. After the advent of screening programs for gastric cancer in South Korea and Japan, up to 50–70% of cases with gastric cancers have been diagnosed at an early stage[2-4]. With the increasing rate of diagnosis at early stages, endoscopic submucosal dissection (ESD) is being actively applied for the minimally invasive treatment of early gastric cancer (EGC) without suspicion of regional lymph node metastasis[5, 6].

In accordance with the current trend of active use of ESD, it is necessary to pay attention to the post-ESD complications. Bleeding is one of the significant complications, with an incidence of 3.6–6.9%[7, 8]. Because bleeding after ESD requires hospitalization and hemostatic interventions, there is a need to predict patients at a high risk of bleeding after ESD. Therefore, there have been reports on risk factors related to bleeding after ESD[9-16]. Recently, a predictive risk-scoring model for bleeding after ESD was proposed in Japan; this tool is expected to raise awareness regarding the potential bleeding sources and thus, help physicians manage patients with EGC who are treated with ESD[17].

Currently, artificial intelligence systems are being applied in various fields of gastroenterology[18]. The machine learning models showed good performance in the triage of necessity for intervention in patients with upper gastrointestinal bleeding and predicting recurrent ulcer bleeding [19, 20]. Deep learning is advantageous over the machine learning model among artificial intelligence systems; its performance is optimized by automatic learning while experiencing various cases. It can integrate and interpret multiple factors simultaneously without external intervention. Hence, the automatically trained deep learning model can generalize well. There has been no study on the efficacy of deep learning for predicting post-ESD bleeding (PEB), and no study has compared these systems with a clinical model.

This study aimed to develop and compare the performance of the deep learning and clinical model for predicting PEB in EGC patients. We chose deep learning among the artificial intelligence systems as a sophisticated algorithm.

MATERIALS AND METHODS

Patients

Patients who underwent ESD for EGC between January 2010 and June 2020 at the Samsung Medical Center, Seoul, South Korea, were screened retrospectively. We excluded cases with: failure to complete ESD (n = 1); prior gastrectomy (n = 2); additional gastrectomy within 28 days after ESD (n = 497); no residual tumor in the ESD specimen (n = 48); multiple procedures, such as EMR for other benign lesions and ESD for EGC (n = 46); and missing values for important variables (n = 7) (Figure 1). A total of 5,629 patients were included in the analysis, who were randomly categorized into the development set (80%) and the validation set (20%). The Institutional review board of the Samsung Medical Center, Korea, approved this study, and the requirement for obtaining informed consent was waived owing to the study's retrospective nature.

Outcome, data sources, study variables, and definitions

The main outcome included the development of a deep learning model and a clinical model that predict the bleeding after ESD in patients with EGC and the comparison of performance between the deep learning model and the clinical model.

The variables used to build the deep learning and clinical models were collected from the medical records retrospectively based on the date of ESD. These variables included: age; sex; comorbidities such as hypertension, diabetes mellitus, liver cirrhosis, and chronic kidney disease (estimated glomerular filtration rate < 60 mL/min/1.73 m^2); patient management with antithrombotic agents [ATs; aspirin, P2Y12RA, warfarin, direct-acting oral anticoagulants (DOAC), and cilostazol], non-steroidal anti-inflammatory drugs (NSAIDs), interruption of ATs, replacement of antiplatelet agents (APA), and heparin bridging; tumor characteristics (single or multiple lesions, location, pathologic size, type of differentiation); piecemeal resection; and laboratory data (albumin level and international normalized ratio [INR]).

Bleeding after ESD was defined as the presence of signs of bleeding (melena, hematemesis, or a decrease in the hemoglobin level by > 2 g/dL), along with endoscopic stigmata of recent bleeding, such as Forrest class Ia, Ib, IIa, and IIb, within 28 days after ESD. Interruption of ATs was defined as the discontinuation of these medications before the procedure, according to the recommended duration. Replacement of APA was described as when the procedure was performed with aspirin or cilostazol alone in patients who were receiving multiple APAs. Heparin bridging was defined as the administration of heparin during the period between the discontinuation and resumption of anticoagulants. A hemoglobin reduction of >2 g/dL was evaluated by calculating the differences in the hemoglobin levels between the day before and after ESD.

Development of the deep learning and clinical models

We built a deep learning model and a clinical model based on the development set, which comprised 80% of the overall cohort. Subsequently, we validated the deep learning and clinical models in the validation set, which comprised 20% of the overall cohort. The categorical variables were converted using one-hot encoding and the continuous variables were normalized, as preprocessing. We built the deep learning model as follows: First, we augmented the development set using the borderline synthetic minority over-sampling technique to overcome the imbalance of the dataset. Synthetic data were generated from 5–100% of the majority class. Second, we constructed the deep learning model using automated machine learning, called Keras Tuner, to tune hyperparameters automatically. The initial architecture of the model was configured similarly to a transformer based on the attention mechanism[21]. Then, we set the number of neurons as a hyperparameter variable, ranging from 12 to 24, in 4 dense layers. The learning rate was also set to a range from 1e-2 − to 1e-4 −. The combination of hyperparameters was determined using Bayesian optimization. Finally, we evaluated the performance in the validation set using a model tuned with the 20% of synthetic data of the majority class. The optimal units of dense layers were selected to 24. The optimal number of attention head was chosen to 16. The architecture is depicted in Supplementary Figure 1. The optimal learning rate with Adam optimizer was 1e-3.

Multivariable logistic regression analysis was performed in the development set to build the clinical model. Then, the clinical model was constructed as a formula with the sum of the beta coefficient values of significant factors with a p-value of <0.05.

The calculated value from the deep learning and clinical models was multiplied by 1,000 and converted as a score. The score that indicated the risk probability was divided by the decile in the development set. We selected cutoff to discriminate the risk categories as low-, intermediate-, and high-risk at a bleeding rate of <5% and <9% in the development set referred to in a previous report[17]. Decile 1st to 4th was allocated to low risk, 5th to 8th to intermediate risk, 9th to 10th to high-risk category. Link to the deep learning and clinical models: https:// github.com/YeongChanLee/Predict-PEB.

Statistical analysis

Descriptive statistics for continuous and categorical variables are presented as means (standard deviation) and frequencies (%). The deep learning model and the clinical model for prediction of bleeding after ESD were evaluated using two methods. First, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and receiver operating characteristic area (ROC) curve along with the area under the curve (AUC) were analyzed. The performance with AUC was compared using the bootstrap test. Second, the risk stratification of PEB based on the development set was applied to the validation set and compared with the actual bleeding rate in the validation set. For example, if the score of calculated cases belongs to the high-risk category, we verified that the real bleeding rate was in the predicted range of nine percent or higher. The predictors for PEB were identified with multivariable logistic regression analysis in the entire cohort and development set. Model development for deep learning was performed using Tensor Flow 2.4.0, and Python 3.8.5. Statistical analyses were performed using the R software (version 3.5.1, Vienna, Austria).

RESULTS

Baseline characteristics

Of the 5,629 patients, 325 experienced post-ESD bleeding (PEB). The non-PEB and PEB groups were comparable in age, liver cirrhosis status, albumin level, INR level, a proportion of aspirin or cilostazol use, undifferentiated tumor type, and piecemeal resection. The PEB group had a higher proportion of males and comorbidities (hypertension, diabetes mellitus, and chronic kidney disease) than the non-PEB group. P2Y12RA and anticoagulants (warfarin or DOAC) and the proportion of patients receiving replacement therapy or heparin bridging were higher in the PEB group than in the non-PEB group. The PEB group had a higher proportion of multiple tumors, middle location of tumors, and larger size of tumors than the non-PEB group. (Table 1). There was no difference in the baseline characteristics between the development and validation sets (Supplementary Table 1).

Predictors for bleeding after ESD

In the overall cohort, the independent predictors were identified as follows: age [OR, 0.98; 95% confidence interval (CI), 0.96–0.99; p value, <0.001], male (OR, 1.65; 95%CI, 1.19–2.28; p value, 0.003), hypertension (OR, 1.56; 95%CI, 1.19–2.03; p value, 0.001), chronic kidney disease (OR, 1.78; 95%CI, 1.18–2.70; p value, 0.006), P2Y12RA (OR, 2.40; 95%CI, 1.22–4.74; p value, 0.011), DOAC (OR, 4.31; 95%CI, 1.26–14.78; p value, 0.020), middle location (OR, 1.72; 95%CI, 1.07–2.74; p value, 0.024), and size (OR, 1.03; 95%CI, 1.02–1.04; p value, <0.001) (Supplementary Table 2).

In the development set, age (OR, 0.98; 95%CI, 0.96–0.99; p value, 0.001), male (OR, 1.54; 95%CI, 1.09–2.19; p value, 0.015), hypertension (OR, 1.35; 95%CI, 1.00–1.82; p value, 0.049), chronic kidney disease (OR, 1.78; 95%CI, 1.12–2.84; p value, 0.015), P2Y12RA (OR, 2.26; 95%CI, 1.05–4.88; p value, 0.037), middle location (OR, 1.97; 95%CI, 1.14–3.41; p value, 0.015), and size (OR, 1.04; 95%CI, 1.03–1.05; p value, <0.001) were identified as independent predictors. The clinical model was a formula described bottom of Table 2.

Performance and comparison of deep learning model and clinical model

The deep learning model was found to have a sensitivity of 64.3%, specificity of 74.0%, PPV of 11.4%, NPV of 97.5%, and AUC of 0.71 (95% CI 0.63–0.78). The clinical model had a sensitivity of 69.6%, specificity of 71.0%, PPV of 11.1%, NPV of 97.8%, and AUC of 0.70 (95%CI, 0.62–0.77) (Table 3 and Figure 2). There were no significant differences in the AUCs between the deep learning and clinical models (Table 3).

The score multiplied by 1,000 to the derived value based on the deep learning and clinical models reflects the risk probability and was divided into deciles. The maximum cutoff was 35.9 in low risk, 57.5 in intermediate risk, and over the 57.5 was assigned to a high-risk category of the deep learning model based on development set (Table 4). In the clinical model, the maximum cutoff was 12.7 in low risk, 24.6 in intermediate risk, and over the 24.6 was considered to a high-risk category based on development set (Table 4). In the validate set, the deep learning model showed an actual bleeding rate of 2.2%, 3.9%, and 11.6%; the clinical model showed an actual bleeding rate of 4.0%, 8.8%, and 18.2% in low-, intermediate-, high-risk categories, respectively (Table 4).

DISCUSSION

The deep learning and clinical models for predicting bleeding after ESD in patients with EGC showed good performance. We demonstrated that deep learning and clinical models could stratify the PEB risk, which correlated with actual bleeding rates. Hence, we suggest that the deep learning model can aid in the prediction of bleeding after ESD, in addition to the clinical model.

This study was the first to establish a deep learning model for predicting bleeding after ESD and demonstrate its performance compared to that of a clinical model. The strengths of this study were its large sample size and the relatively recent data from a single institution. In addition, we included all essential variables and sought the advantages of the deep learning model that can deal with extensive data, complex problems and improve its performance incrementally by automated learning. We included all types of ATs separately and clarified the distinction between patients without an indication for ATs, patients who received an interruption before the procedure, and patients who received replacement or heparin bridging.

Our study identified younger age, male sex, hypertension, chronic kidney disease, P2Y12RA use, DOAC use, middle tumor location, and tumor size as the predictors of PEB. Previous studies also reported that younger age was associated with PEB[11, 22, 23]. It is unclear why younger age was associated with PEB. Several reports proposed that atrophic change along with aging might relate to decreasing the vascularity on the mucosal and submucosal layers[11, 22, 24-26]. Although aging and changes in intestinal vasculature have not been clearly elucidated, a decrease in the volume of vasculature with aging was observed in animals[27]. Aspirin did not increase the PEB risk after discontinuation about one week[13]. Although some reported that maintaining aspirin did not increase the PEB risk[10, 13, 28, 29], a meta-analysis showed that aspirin was associated with increased bleeding risk, requiring clinical caution[30]. There is still controversial due to limited evidence for P2Y12R [9, 11, 17, 31]. In comparision, an increased bleeding risk after ESD has been reported consistently in patients receiving dual antiplatelets. In addition, there were reports that warfarin or DOAC are related to bleeding risk [17]; rather, some reported heparin bridging was associated with PEB risk [9, 10]. The irony is that most of the patients who experience heparin bridging take warfarin or DOAC, but the results about each factor were inconsistent in previous retrospective studies. It is assumed that the duration of discontinuation and other individual factors might influence these results. In addition, it has been suggested that large size[8, 11, 23], CKD with hemodialysis [10, 17, 32], and long procedure time [11] were associated with bleeding after ESD. The upper location showed increased PEB risk [22, 33], contrarily some other reported lower location related to increased PEB risk [22, 33]; a recent meta-analysis did not prove significance according to the location[8]. Recently, a predictive risk-scoring model for PEB in Japan showed that chronic kidney disease with hemodialysis, usage of aspirin, P2Y12RA, cilostazol, warfarin, DOAC, lower third tumor location, tumor size > 30 mm, and the presence of multiple tumors were the predictors of PEB, whereas interruption was a protective factor against PEB[17]. Another recent model proposed a simple algorithm including significant factors with continuous use of ATs, size ≥ 49 mm, age <62 years. We also found an association between P2Y12RA or DOAC usage and PEB; however, other ATs were not associated with PEB, and interruption and heparin bridging or replacement of APA were not identified as the protective factors. In our institution, ESD is classified as a high-risk procedure based on the national practice guidelines, and experts are consulted before ESD in patients receiving ATs. The expert assesses the thromboembolic risk depending on the underlying disease and recommends the possibility of interruption, duration of interruption, and the need for heparin bridging or replacement of APA[34-37]. Recently, a guideline published in South Korea also categorized ESD as an ultra-high-risk procedure and recommended interruption of ATs with heparin bridging or replacement of APA according to the thromboembolic risk[38].

The deep learning model in our study showed an AUC of 0.71, which was comparable to the AUC of 0.72 for a risk-scoring model in Japan[17] and the AUC of 0.70 for the clinical model in our study. In the validation set, predicted as low-, intermediate-, high-risk categories showed an actual bleeding rate of 2.2%, 3.9%, and 11.6% in the deep learning; 4.0%, 8.8%, and 18.2% in the clinical model. Our study demonstrated that the deep learning and clinical models can stratify the bleeding risk after ESD. The predicted risk categories correlated with actual bleeding rate; even considering the actual bleeding rate was slightly lower than predicted range of ≥5% and <9% (intermediate risk) in the deep learning and was close to upper range in the clinical model. Our findings support the clinical potential of the deep learning model for predicting PEB risk based on its comparable performance. Because bleeding after ESD requires intervention and hospitalization, physicians are concerned about the occurrence of PEB as a major complication. Based on the risk-prediction model, physicians could carefully assess the bleeding risk and perform preventive hemostasis during the procedure. Suppose additional management like the shielding method for preventing PEB in the selected high-risk group is attempted. In that case, it is anticipated that the deep learning model could support risk stratification.

Our study has several limitations. Due to its retrospective design, information such as the timing of the resumption of ATs, endoscopist's experience, defect size, and procedure duration was missing. Furthermore, our study was designed as a single-center study; hence, hospital-based validation in other hospitals was not performed, and further proof is warranted. However, the deep learning model might be generalizable because it automatically identifies the risk or probability of bleeding without the external intervention of known relevant factors. Both the deep learning and clinical models showed a low PPV, which may be related to the low incidence of bleeding after ESD, even though bleeding is one of the major complications. In our cohort, the number of patients who received anticoagulants (warfarin or DOAC) was small; therefore, it is possible that the statistical significance of these variables was insufficient for establishing a clinical model in the development set. In this regard, despite the fact that our study focused on the development of a deep learning model and a clinical model, as well as the utility of the deep learning model, further accumulation of data and additional analysis will be required before the commencement of the clinical application of artificial intelligence systems.

CONCLUSION

In conclusion, we introduced a deep learning model to predict the risk of bleeding after ESD in patients with EGC. The model demonstrated its performance as comparable to the clinical model. The deep learning model could help the physicians raise caution to the PEB and would be a desirable tool for supporting ESD application.

72837_Auto_Edited.docx

ORIGINALITY REPORT

8%

SIMILARITY INDEX

PRIMARY SOURCES

1	www.mdpi.com Internet	71 words —	2%
2	cris.bgu.ac.il Internet	64 words —	2%
3	coek.info Internet	36 words —	1%
4	www.ncbi.nlm.nih.gov Internet	36 words —	1%
5	www.thieme-connect.com Internet	35 words —	1%
6	journals.sagepub.com Internet	20 words —	1%
7	www.science.gov Internet	19 words —	1%

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	< 1%
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 10 WORDS