A deep learning algorithm to detect anaemia with ECGs: a retrospective, multicentre study



Joon-myoung Kwon*, Younghoon Cho*, Ki-Hyun Jeon, Soohyun Cho, Kyung-Hee Kim, Seung Don Baek, Soomin Jeung, Jinsik Park, Byung-Hee Oh

Oa OPEN ACCESS

Summary

Background Anaemia is an important health-care burden globally, and screening for anaemia is crucial to prevent multi-organ injury, irreversible complications, and life-threatening adverse events. We aimed to establish whether a deep learning algorithm (DLA) that enables non-invasive anaemia screening from electrocardiograms (ECGs) might improve the detection of anaemia.

Methods We did a retrospective, multicentre, diagnostic study in which a DLA was developed using ECGs and then internally and externally validated. We used data from two hospitals, Sejong General Hospital (hospital A) and Mediplex Sejong Hospital (hospital B), in South Korea. Data from hospital A was for DLA development and internal validation, and data from hospital B was for external validation. We included individuals who had at least one ECG with a haemoglobin measurement within 1 h of the index ECG and excluded individuals with missing demographic, electrocardiographic, or haemoglobin information. Three types of DLA were developed with 12-lead, 6-lead (limb lead), and single-lead (lead I) ECGs to detect haemoglobin concentrations of 10 g/dL or less. The DLA was built by a convolutional neural network and used 500-Hz raw ECG, age, and sex as input data.

Findings The study period ran from Oct 1, 2016, to Sept 30, 2019, in hospital A and March 1, 2017, to Sept 30, 2019, in hospital B. 40 513 patients at hospital A and 4737 patients at hospital B were eligible for inclusion. We excluded 281 patients at hospital A and 72 patients at hospital B because of missing values for clinical information and ECG data. The development dataset comprised 57 435 ECGs from 31898 patients, and the algorithm was internally validated with 7974 ECGs from 7974 patients. The external validation dataset included 4665 ECGs from 4665 patients. 586 (internal) and 194 (external) patients within the combined dataset were found to be anaemic. During internal and external validation, the area under the receiver operating characteristics curve (AUROC) of the DLA using a 12-lead ECG for detecting anaemia was 0.923 for internal validation and 0.901 for external validation. Using a 90% sensitivity operating point for the development data, the sensitivity, specificity, negative predictive value, and positive predictive value of internal validation were 89.8%, 81.5%, 99.4%, and 20.0%, respectively, and those of external validation were 86.1%, 76.2%, 99.2%, and 13.5%, respectively. The DLA focused on the QRS complex for deciding the presence of anaemia in a sensitivity map. The AUROCs of DLAs using 6 leads and a single lead were in the range of 0.841-0.890.

Interpretation In this study, using raw ECG data, a DLA accurately detected anaemia. The application of artificial intelligence to ECGs could enable screening for anaemia.

Funding None.

Copyright © 2020 The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY-NC-ND 4.0 license.

Introduction

Anaemia, defined as reduced haemoglobin concentration, is the most common haematological disorder, affecting more than 2 billion people globally. It can be caused by various acute and chronic diseases, including gastrointestinal bleeding, malnutrition, malignancy, and chronic kidney disease. Many causes of anaemia can be managed with simple treatment, whereas other causes such as thalassaemia need frequent monitoring. Some emergent causes of anaemia, such as major gastrointestinal bleeding, are life-threatening and require early detection.³

As the symptoms of anaemia are vague, it is difficult to diagnose with only patient histories and exams until the condition is uncompensated and complications occur.^{3,4} The gold standard of diagnosing anaemia is

a laboratory exam to measure the concentration of haemoglobin via a complete blood count. The laboratory test is invasive, costly, and requires specialised equipment and infrastructure (eg, trained medical staff for sampling blood and a haematology analyser for assessment with a biochemical reagent.⁵

The physiological role of haemoglobin is the delivery of oxygen to vital organs, including the heart. Anaemia is a well known cause of ischaemic cardiovascular diseases and heart failure, and is also associated with increased cardiovascular mortality.⁶⁻⁸ In several previous studies, anaemia was shown to change the morphology of the electrocardiogram (ECG), and researchers suggested that mismatching oxygen demand and supply in the myocardium affects the ECG.⁹⁻¹² However, it is not easy to

Lancet Digital Health 2020; 2: e358-67

See Comment page e334
*Contributed equally

Department of Emergency Medicine (J-M Kwon MD), Division of Cardiology, Cardiovascular Center (K-H Jeon MD, K-H Kim MD, J Park MD, B-H Oh MD), and Division of Nephrology, Department of Internal Medicine (S D Baek MD,

S Jeung MD), Mediplex Sejong Hospital, Incheon, South Korea; Artificial Intelligence and Big Data Research Center, Sejong Medical Research Institute, Bucheon, South Korea (J-M Kwon, K-H Kim, K-H Jeon); Medical research team, Medical AI, Seoul, South Korea (J-M Kwon, J Park); and Medical Research and Development Center, Bodyfriend, Seoul,

Correspondence to:
Dr Ki-Hyun Jeon, Division of
Cardiology, Department of
Internal Medicine, Cardiovascular
Center, Mediplex Sejong
Hospital, Incheon 21080,
South Korea
imcardio@gmail.com

South Korea (Y Cho MD,

S Cho MD)

Research in context

Evidence before this study

Anaemia is a well known cause of ischaemic cardiovascular disease and heart failure and is also associated with increased cardiovascular mortality. Because haemoglobin is important in oxygen delivery to vital organs, anaemia can cause a mismatch between the demand and supply of oxygen in the myocardium. We searched PubMed for algorithms to detect anaemia with electrocardiograms (ECGs) from database inception up to Dec 11, 2019, using the terms "anemia" AND "electrocardiogram" AND/OR "deep learning" OR "machine learning" OR "artificial intelligence". Previous studies showed that several electrocardiographic changes can be seen in patients with severe anaemia, including a small QRS amplitude, prolonged QT interval, T-wave inversion, rightward T-wave axis deviation, and minor atrioventricular conduction disturbance. However, to our knowledge, no study has been published that developed and validated a deep learning algorithm for detecting anaemia using ECGs.

Added value of this study

This is the first study to develop and validate a deep learning based artificial intelligence algorithm for detecting anaemia using ECGs. The researchers developed an artificial intelligence algorithm for detecting clinical information from the ECG. Artificial intelligence can diagnose heart failure and predict atrial fibrillation during sinus rhythm, diverse conditions that clinicians cannot diagnose with a single ECG. The diagnostic

performances of artificial intelligence algorithms increase in accuracy for detecting left ventricular hypertrophy, mitral regurgitation, and hyperkalaemia by ECG. In this study, the algorithm showed high performance for screening anaemia and revealed that a deep learning algorithm can identify subtle ECG changes in patients with anaemia. This study can expand artificial intelligence ECG knowledge from heart disease to systemic disease and we believe that it can be hypothesisgenerating research to understand the subtle relationship between electrophysiology and cardiac or systemic diseases.

Implications of all the available evidence

Broadly, the use of artificial intelligence, based on a deep learning algorithm, has been enabled by use of labelled big data, along with markedly enhanced computing power and cloud storage in the medical field. These innovations will transition medicine in the future into so-called 4P medicine (prevention, prediction, personalisation, and participation [or precision]). This study could become a tool for this innovation in the medical field. The algorithm for this study can detect subtle ECG changes along with decreased haemoglobin using single-lead ECG as well as 12-lead ECG. It can be applied to detect anaemia in the hospital setting, and could also potentially be used in wearable devices for remote monitoring of patients at risk, such as those with chronic kidney disease and congestive heart failure.

detect such subtle ECG changes, so the current state of the ECG is not useful for detecting anaemia. Screening for anaemia with an ECG would be straightforward, and patients who were suspected to be anaemic could be referred for confirmatory laboratory tests.

In this study, we aimed to develop and validate a deep learning algorithm (DLA) for detecting anaemia with ECGs. To develop a reliable anaemia screening method based on the ECG, we used deep learning, which has previously been used to identify lesions in medical images such as retinal images and chest radiographies.¹³ In the past 3 years, deep learning has been used to analyse ECGs to approximate atrial fibrillation, heart failure, and even age and sex.¹⁴⁻¹⁷ We hypothesised that a DLA could effectively screen for anaemia.

Methods

Study design and participants

We did a retrospective, multicentre, diagnostic study in which a DLA was developed using ECGs and then internally and externally validated. We excluded individuals with missing demographic, electrocardiographic, or haemoglobin information. Data from hospital A (Sejong General Hospital, Bucheon, South Korea) were used for development and internal validation. In hospital A, we identified patients with at least one

standard digital, 10-s, 12-lead ECG acquired in the supine position within the study period (Oct 1, 2016, to Sept 30, 2019) and at least one haemoglobin measurement obtained within 1 h of the index ECG. The individuals who visited the outpatient department and emergency department and were admitted to hospital A were the study population for the development and internal validation datasets. Patients treated at hospital A were randomly split into algorithm development (80%) and internal validation (20%) datasets, exclusively. Data from hospital B (Mediplex Sejong Hospital, Incheon, South Korea) were used for external validation. We identified the patients who were admitted to hospital B in the study period (March 1, 2017, to Sept 30, 2019) and who had at least one ECG and one haemoglobin measurement obtained within 1 h of the index ECG. Because the purpose of the validation data was to assess the accuracy of the algorithm, we only used one ECG from each patient for the internal and external validation datasets the haemoglobin test closest to their most recent ECG in the study period.

Although Sejong General Hospital and Mediplex Sejong Hospital belong to the same hospital group, they are separate. Sejong General Hospital opened in 1982, is located in Bucheon, and is a cardiology and heart surgery specialty hospital. More than 50% of the diseases covered

by the hospital are cardiovascular. By contrast, Mediplex Sejong Hospital opened in 2017 as a community hospital in Incheon. The proportion of patients with cardiovascular disease in Mediplex Sejong Hospital is about 20-30%. The distance between the two hospitals is about 10 miles, but the patient groups of the two hospitals do not overlap (as shown by patient records) due to geographical factors and public transport connections. The doctors and other medical staff are hired separately in each hospital. The ECG machines in the two hospitals were manufactured by different companies and those of Sejong General Hospital were older machines.

This study was approved by the institutional review boards of Sejong General Hospital (2019-0579, Bucheon, South Korea) and Mediplex Sejong Hospital (2019-133, Incheon, South Korea). Clinical data, including digitally stored ECGs, haemoglobin values, age, and sex, came from both hospitals. Both institutional review boards waived the need for informed consent because of the retrospective nature of the study using fully anonymised ECG and health data, and minimal harm.

Procedures

Predictor variables were ECG, age, and sex. Digitally stored 12-lead ECG data, amounting to 5000 numbers for each lead, were recorded over 10 s (500 Hz). We removed 1 s each at the beginning and end of each ECG because these areas have more artifacts than other parts. Because of this, the length of each ECG was 8 s (4000 numbers). We made a dataset using the entire 12-lead ECG data. We also used partial datasets from 12-lead ECG data, such as limb 6-lead (aVL, I, -aVR, II, aVF, and III), and single lead (I). We selected the sets of leads because these leads can easily be recorded by wearable and pad devices in contact with the hands and legs. 18 Consequently, when we developed and validated an algorithm using 12-lead ECGs, we used the dataset that was 2-dimensional (2D) data of 12×4000 numbers. To make the input 2D ECG data, we rearranged the data in the order of V1, V2, V3, V4, V6, aVL, I, -aVR, II, aVF, and III. The convolutional neural network (CNN), a method of deep learning, is a well known method in computer science for learning 2D image data.13 In the same manner, when we developed and validated an algorithm using 6-lead ECGs, we used datasets that were 6×4000, and when using single-lead ECGs, we used datasets that were 1×4000 .

The endpoint of this research was moderate to severe anaemia, defined as a serum haemoglobin concentration of 10.0 g/dL or less.19

The DLA was developed using many hidden layers of neurons to learn complex hierarchical non-linear representations from the data.13 As a block with six stages, there were two convolution layers, two batch normalisations, one max pooling, and one dropout layer repeated (appendix p 1). Although dropout was usually applied to the fully connected layer, not to the convolutional layer, the dropout in the convolutional layer built a robust algorithm and has shown enhanced performance in several studies.20 We used 4×1 max-pooling layers between blocks 1 and 4, and 4×2 max-pooling layers between blocks 4 and 6. As the shape of the input data was an asymmetrical rectangle, the initial max-pooling was used only for pooling long (time-axis) lengths. The last convolutional layer of the CNN connected to a flattened layer, which was fully connected to the 1-dimensional (1D) layer composed of 128 nodes. The input layer of epidemiology data (age and sex) was concatenated with the 1D layer. There are two fully connected 1D layers after the flattened layer, and the second fully connected 1D layer was connected to the output node, which was composed of one node. The output values of the output node represented the possibility of anaemia, and the output node used a sigmoid function as an activation function because the output of the sigmoid function was between 0 and 1. The training process is provided in the appendix (p 1). We did additional analyses for DLA using limb 6-lead and single-lead ECGs. The architecture of a DLA for limb 6-lead and single-lead ECGs is provided in the appendix (pp 3–6).

For comparing the performances of the DLAs, we developed several DLAs using other combinations of data. We developed a DLA using raw data from a 12-lead ECG, without age and sex; using features of an ECG (heart rate, presence of atrial fibrillation or flutter, PR interval, QRS duration, QT interval, QTc, P-wave axis, R-wave axis, and T-wave axis) with age and sex; and finally using features of an ECG only. We used a multilayer perceptron to develop the DLA using features of the ECG. Furthermore, to confirm the additional value of use of deep learning, we developed logistic regression models using ECG features or ECG features plus age and

Because this was a retrospective study, the ECGs had been evaluated according to clinical need (an ECG costs less than US\$10 in South Korea, so we routinely check ECGs in various clinical situations). In some cases, the same clinicians ordered ECG and complete blood count, but in other cases different clinicians ordered them.

Statistical analysis

Continuous variables are presented as mean (SD) and were compared using the unpaired Student's t test or Mann-Whitney *U* test. Categorical variables are expressed as frequencies and percentages and compared using the χ^2 test.

At each input (ECG, age, and sex) of validation data, the DLA calculated the possibility of clinically significant anaemia in the range from 0 (non-anaemia) to 1 (anaemia). To confirm the performance of the DLA, we compared the possibility calculated by the DLA with the presence of anaemia in the internal and external validation datasets. For this, we used the area under the See Online for appendix receiver operating characteristics curve (AUROC) and the area under the precision-recall curve (AUPRC) to measure the performance of the model.²¹ Exact 95% CIs

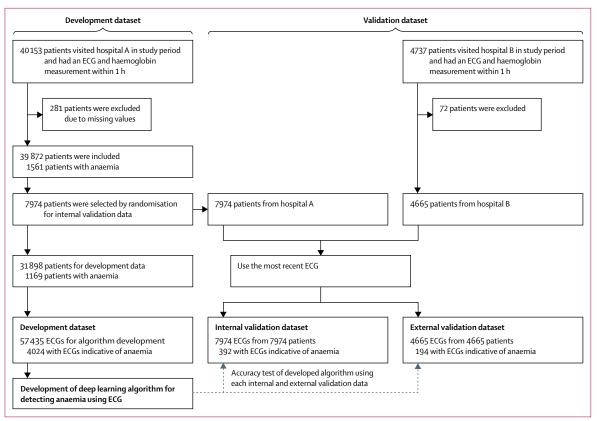


Figure 1: Study profile ECG=electrocardiogram.

were used for all measures of diagnostic performance except for AUROC and AUPRC. The CIs for AUROC and AUPRC were determined based on Sun and Su optimisation of the De-long method using the pROC package in R (The R Foundation, Vienna, Austria). A significant difference in patients' characteristics was defined as a two-sided p value of less than 0.001.

We confirmed the performance of the DLA in a subanalysis of age, sex, body-mass index, reason for hospital admission (acute disease or chronic disease), and admission department (non-surgery or surgery department) using external validation data. For confirming the performance of the DLA without the effect of other variables, we used a multivariable logistical analysis with the predictive value of the DLA using 12-lead ECG raw data, age, sex, body-mass index, heart rate, and presence of atrial fibrillation and flutter.

We confirmed the Pearson's correlation between predicted haemoglobin concentrations from the DLA and actual haemoglobin concentrations in the internal and external validation data. For this analysis, we used transfer learning methods. We reused the developed DLA and changed the activation function of the last node from sigmoid to linear. We did a subgroup analysis of patients who were in the internal and external validation datasets and were given a red blood cell transfusion. We

confirmed the predicted scores from the DLA before and after transfusion.

To understand the developed DLA and make a comparison with existing medical knowledge, it was important to identify which ECG region had a substantial effect on the decision of the DLA. We used a sensitivity map with a saliency method to visualise the ECG region used by the DLA to detect anaemia. The map was computed with the first-order gradients of the classifier probabilities with respect to the input signals. If the probability of a classifier was sensitive to a specific region of the signal, the region would be considered as important in the model. We used a gradient-weighted class activation map for visualisation. 22 Such an activation map uses the gradient information of the algorithm and can be used in any activation function and any architecture of CNN. Statistical analyses were computed using R software, version 3.4.2. We also used TensorFlow's open-source software library as the backend, and Python (version 3.5.2) for the analysis.

Role of the funding source

No funding was secured for this study. Bodyfriend provided support in the form of salaries for the authors (YC, SC), but had no role in study design, data collection, data analysis, decision to publish, or preparation of the

	All participants (n=31 898)	Normal to mild anaemia (n=30729)	Moderate anaemia (n=1148)	Severe anaemia (n=21)	p value
Sex					<0.0001
Women	15 345 (48-1%)	14571 (47-4%)	764 (66-6%)	10 (47-6%)	
Men	16 553 (51.9%)	16 158 (52-6%)	384 (33-4%)	11 (52-4%)	
Age, years	59.70 (16.45)	59-29 (16-32)	70.66 (16.08)	72·10 (12·19)	<0.0001
Body-mass index, kg/m²	24.70 (4.11)	24-75 (4-12)	23.56 (3.75)	25-41 (4-25)	<0.0001
Haemoglobin, g/dL	13.59 (1.88)	13-77 (1-64)	8.82 (1.06)	4.21 (0.63)	<0.0001
Heart rate, bpm	72.80 (18.44)	72-51 (18-26)	80.45 (21.30)	90.05 (18.14)	<0.0001
PR interval, ms	171-24 (29-87)	171-13 (29-51)	174-50 (38-69)	157-25 (51-94)	0.0012
P-wave axis	43.72 (29.72)	43.76 (29.41)	42-31 (37-89)	59·14 (47·77)	0.058
R-wave axis	39-61 (44-28)	39-67 (44-08)	38.00 (49.45)	46-33 (47-24)	0.36
T-wave axis	45.22 (48.61)	44-57 (47-53)	62-35 (69-63)	75.19 (68.02)	<0.0001
QT inverval, ms	404-43 (42-42)	404-43 (41-94)	404-74 (53-67)	388-48 (36-54)	0.22
QTc, ms	438.04 (33.70)	437-27 (33-30)	457-96 (37-95)	470-14 (33-51)	<0.0001
QRS duration, ms	95.66 (17.50)	95.60 (17.28)	97:12 (22:56)	101-19 (22-18)	0.0051
QRS amplitude, mV	1.54 (0.64)	1.53 (0.63)	1.65 (0.89)	1.37 (0.53)	<0.0001
ECG, number per patient	57 435 (1.80)	53 411 (1.74)	3930 (3·42)	94 (4·48)	

Data are n (%) or mean (SD), unless specified. Normal to mild anaemia (haemoglobin ≥10 g/dL), moderate anaemia (haemoglobin ≥6-10 g/dL), and severe anaemia

manuscript. The corresponding author had full access to all anonymised datasets and summary estimates from each dataset, and had final responsibility for the decision

Table 1: Baseline characteristics for the development dataset

(haemoglobin <6 g/dL). ECG=electrocardiogram.

to submit for publication.

Results

The study period ran from Oct 1, 2016, to Sept 30, 2019, in hospital A and March 1, 2017, to Sept 30, 2019, in hospital B. 40 513 patients at hospital A and 4737 patients at hospital B were eligible for inclusion. We excluded 281 patients at hospital A and 72 patients at hospital B because of missing values for clinical information and ECG data. The development dataset from hospital A included 57435 ECGs of 31898 patients, of whom 1169 had moderate to severe anaemia, as shown in figure 1 and table 1. ECGs indicative of anaemia were more likely to be recorded in patients who were older, as shown in table 1. In severe anaemia, defined as haemoglobin less than 6 g/dL, the ECGs had lower amplitude QRS, prolonged QRS duration, prolonged corrected QT interval, rightward T-wave axis, prolonged PR interval, and tachycardia (table 1). The performance of the algorithm was then confirmed using 7974 ECGs from 7974 patients in the internal validation data from hospital A, and 4665 ECGs from 4665 patients in the external validation data from hospital B. Internal validation data had 392 (4.9%) patients with clinically significant anaemia and external validation data had 194 (4.2%) patients with clinically significant anaemia. The two validation datasets had significantly different characteristics, as shown in tables 1, 2.

For the endpoint moderate to severe anaemia, the AUROC of the 12-lead DLA was 0.923 (95% CI

	All (n=12639)	Hospital A (n=7974)	Hospital B (n=4665)	p value
Sex				0.0008
Women	6114 (48-4%)	3854 (48-3%)	2260 (48-4%)	
Men	6525 (51-6%)	4120 (51-7%)	2405 (51-6%)	
Age, years	58-56 (16-90)	59-19 (16-58)	57.48 (17.37)	<0.0001
Body-mass index, kg/m²	24.65 (3.77)	24-64 (3-77)	24-67 (3-77)	0.75
Haemoglobin, g/dL	13.54 (1.92)	13-57 (1-92)	13-49 (1-91)	0.024
Heart rate, bpm	74.00 (18.92)	72.79 (18.88)	76.06 (18.82)	<0.0001
PR interval, ms	169-26 (28-98)	170-75 (29-40)	166-80 (28-11)	<0.0001
P-wave axis	43.82 (29.56)	43-66 (30-06)	44.08 (28.72)	0.47
R-wave axis	38-83 (44-31)	39-21 (45-21)	38-19 (42-74)	0.21
T-wave axis	43.95 (47.78)	44-49 (48-14)	43.02 (47.15)	0.096
QT inverval, ms	401-95 (43-66)	404.79 (43.27)	397-10 (43-91)	<0.0001
QTc, ms	438-35 (35-76)	438-00 (33-33)	438-95 (39-57)	0.15
QRS duration, ms	95-99 (17-49)	95.75 (17.67)	96.42 (17.18)	0.037
QRS amplitude, mV	1.56 (0.74)	1.54 (0.71)	1.59 (0.78)	<0.0001
Anaemia	586 (4.6%)	392 (4.9%)	194 (4·2%)	0.056
ECG, number per patient	12 639 (1.0)	7974 (1.0)	4665 (1.0)	

Data are n (%) or mean (SD), unless specified. Hospital A was a cardiovascular teaching hospital with all patients (admission, outpatient clinic, emergency department). Hospital B was a community general hospital with admitted patients. Hospital A was for internal validation, and hospital B for external validation. ECG=electrocardiogram.

Table 2: Baseline characteristics for internal and external validation datasets

0.909-0.937) and the AUPRC of the 12-lead DLA was 0.454 (0.450-0.458) during internal validation. In external validation for the endpoint, the AUROC of the DLA was 0.901 (95% CI 0.881-0.921) and the AUPRC was 0.346 (0.341-0.350). As shown in table 3 and figure 2, the AUROC was 0.3-0.6 lower when using 6-lead or single-lead ECG than when using 12-lead ECG. The DLA

AUROC Ambre AURO Part AURO Part AURO Part Part Part New Part New Part Part Part New Part Par		Internal validation	ıtion						External validation	ion					
13-5%		AUROC (95% CI)	pvalue	AUPRC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)				Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
% 71.1% 11.1% 9-88.0) (697-72.4) (95-12.8) 7-94.4) (46-1-49.1) (6-0-8.1) 8.4 36.2% 5-6% 9-91.5) (34.8-37.6) (4.8-6.5) 8.4 42.1% 6-2% 9-91.2) (40.8-44.0) (5-1-73) 1.% 38.5% 5-9% 1-92.3) (37.0-39.9) (5-0-6.8) 3-83.0) (76.0-78.5) (11.2% 8-83.0) (60.0-78.5) (10.0% 8-92.8) (64.0-66.8) (8.6-11.5)	12-lead ECG raw signal plus age and sex	0.923 (0.909-0.937)	÷	0.454 (0.450-0.458)	89.8% (86.4–92.6)	81.5% (80.6–82.4)	20.0% (18.2–22.0)	99.4% (99·1–99·5)	0.901 (0.881-0.921)	:	0.346 (0.341-0.350)	86·1% (80·4-90·6)	76·2% (74·9-77·4)	13·5% (11·7-15·6)	99.2% (98.9–99.5)
% 47.6% 7.0% -94.4) (46-1-49.1) (6-0-8.1) -91.5) 36.2% 5-6% -91.5) (34.8-37.6) (4.8-6-5) % 42.1% 6-2% -91.2) (40.8-44.0) (5-1-73) % 38-5% 5-9% -92.3) (37-0-39-9) (5-0-6.8) % 77-2% 12-8% 93.3) (7-0-78-5) (11-0-14-9) % 65-4% 10-0% 8-92.8) (64-0-66-8) (8-6-11-5)	12-lead ECG raw signal only	0.883	<0.0001	0·324 (0·321–0·327)	88.0% (84.4-91.1)	70·0% (69·0-71·1)	13·2% (11·9-14·5)	99·1% (98·8-99·4)	0.850 (0.823-0.878)	<0.0001	0.229 (0.225-0.233)	83.0% (76.9-88.0)	71·1% (69·7-72·4)	11·1% (9·5-12·8)	99.0%
\(\psi_{-91.5}\) (34.8-37.6) (4.8-6.5) \(\psi_{-91.2}\) (34.8-37.6) (4.8-6.5) \(\psi_{-91.2}\) (40.8-44.0) (5.1-73) \(\psi_{-92.3}\) (37.0-39.9) (5.0-6.8) \(\psi_{-92.3}\) (37.0-39.9) (5.0-6.8) \(\psi_{-83.0}\) (70.0-78.5) (11.0-14.9) \(\psi_{-92.8}\) (64.0-66.8) (8.6-11.5)	ECG features plus age and sex (deep learning)	0.822 (0.802-0.842)	<0.0001	0.202 (0.199-0.204)		51.4% (50.3–52.6)	8.7%	98.9% (98.6–99.2)	0.800 (0.771-0.830)	<0.0001	0.140 (0.138-0.142)	90.7% (85.7–94.4)	47.6% (46.1–49.1)	7.0% (6.0–8.1)	99.2% (98.7–99.5)
7% 42.1% 6.2% 2-91.2) (40.8-44.0) (5.1-7.3) 1% 38.5% 5.9% 2-92.3) (37.0-39.9) (5.0-6.8) 3% 77.2% 12.8% 5-83.0) (76.0-78.5) (11.0-14.9) 7% 65.4% 10.0% 6-92.8) (64.0-66.8) (8.6-11.5)	ECG features plus age and sex (logistic regression)	0.806 (0.783-0.829)	<0.0001	0.200 (0.197-0.202)	88.3% (84.7–91.3)	46·2% (45·0-47·3)	7.8%	98·7% (98·3-99·0)	0.761 (0.725-0.797)	<0.0001	0·131 (0·128-0·133)	87.1% (81.6–91.5)	36·2% (34·8–37·6)	5.6% (4.8–6.5)	98·5% (97·8–99·0)
1% 38.5% 5.9% -92.3) (37.0-39.9) (5.0-6.8) 3% 77.2% 12.8% 5-83.0) (76.0-78.5) (11.0-14.9) 7% 65.4% 10.0% 5-92.8) (64.0-66.8) (8.6-11.5)	ECG features only (deep learning)	0.767 (0.743-0.790)	<0.0001	0·156 (0·154-0·158)	88.0% (84.4-91.1)	42.7% (41.6–43.8)	7.4% (6.6–8.1)	98.6% (98.1–98.9)	0.764 (0.740–0.779)	<0.0001	0.117 (0.115-0.119)	88.7% (84.2-91.2)	42·1% (40·8–44·0)	6.2% (5.1-7.3)	98.8% (98.3–99.2)
9% 77.2% 12.8% 5-83.0) (76.0-78.5) (11.0-14.9) 7% 65.4% 10.0% 5-92.8) (64.0-66.8) (8.6-11.5)	ECG features only (logistic regression)	0.747 (0.721-0.772)	<0.0001	0·146 (0·142–0·149)	87·2% (83·5-90·4)	40.9% (39.8–42.0)	7·1% (6·4-7·9)	98.4% (97.9–98.8)	0.733 (0.696-0.769)	<0.0001	0.113 (0.111-0.115)	88·1% (82·7-92·3)	38·5% (37·0-39·9)	5.9% (5.0-6.8)	98.7% (98.0–99.2)
% 65.4% 10.0% 5-92.8) (64.0-66.8) (8.6-11.5)	6-lead ECG raw signal plus age and sex	0.890 (0.875-0.905)	<0.0001	0.359 (0.356-0.362)	85.7% (81.9–89.0)	76·3% (75·3-77·3)	15.8% (14·2-17·4)	99.0% (98.8–99.3)	0.848 (0.822-0.874)		0.260 (0.256-0.265)	77-3% (70-8-83-0)	77.2% (76.0–78.5)	12.8% (11.0–14.9)	98·7% (98·3-99·1)
AUROC=area under the receiver operating characteristics curve. AUPRC=area under the precision-recall curve. PPV=positive predictive value. NPV=negative predictive value. ECG=electrocardiogram.	Single-lead ECG raw signal plus age and sex	0.870 (0.853-0.887)	<0.0001	0.316 (0.313-0.319)	87.8% (84·1–90·8)	68.0% (67.0–69.1)	12·4% (11·2-13·7)	99·1% (98·8–99·3)	0.841 (0.815-0.866)		0·185 (0·182-0·189)	88.7% (83.3-92.8)	65.4% (64·0-66·8)	10.0% (8.6–11.5)	99.3% (98.9–99.5)
	AUROC=area under the reco	eiver operating cha	racteristics cu	urve. AUPRC=area u	ınder the precisi	on-recall curve.	. PPV=positive p	predictive value. I	√PV=negative predi	ctive value. E	:CG=electrocardio	gram.			

using 12-lead ECG raw data with age and sex statistically outperformed other DLAs using other combinations of data (ECG raw data only or ECG features). In addition, DLAs outperformed the algorithms that were based on logistic regression. As shown in the appendix (p 6), we found excellent DLA performance across age, sex, body-mass index, reason for admission, and admission department strata. After multivariable logistical regression with age, sex, body-mass index, heart rate, and presence of atrial fibrillation and flutter, we found the predictive values of DLA using the 12-lead ECG raw data had a significant correlation with anaemia (adjusted hazard ratio 127·11, 95% CI 73·84–219·94; p<0·0001) using the internal and external validation data (appendix p 7).

The operating point with high sensitivity (90%) using the development data reflects an output that could be used for a screening tool. The algorithm's sensitivity for validation was $89 \cdot 8\%$ with a specificity of $81 \cdot 5\%$ in internal validation, and a sensitivity of $86 \cdot 1\%$ with a specificity of $76 \cdot 2\%$ in external validation. Given the less than 5% prevalence of anaemia, these findings correspond to a negative predictive value of $99 \cdot 2\%$ (external) to $99 \cdot 4\%$ (internal).

As shown in figure 3, the predicted haemoglobin concentration based on the DLA had a strong correlation with the actual haemoglobin concentration in the internal and external validation data (r=0·854, 95% CI 0·849–0·859, p<0·0001). The mean absolute error between the predicted and actual haemoglobin concentrations was 0·791 (95% CI 0·779–0·802). 57 patients with moderate to severe anaemia were given the appropriate amount of blood in transfusions and the predicted haemoglobin concentration of most patients who received a transfusion increased afterwards in line with an actual increase in haemoglobin concentration (figure 4).

As shown in figure 5, the sensitivity map shows that the DLA focused most on the QRS complex to predict the presence of anaemia. The DLA also focused on T-wave and P-wave data, but the significance of these regions was lower than for the QRS complex.

Discussion

To our knowledge, this is the first study to develop and validate a DLA for detecting anaemia using ECGs. This study reveals that a DLA, a powerful tool of artificial intelligence, can elucidate subtle ECG changes that distinguish patients with anaemia. Many studies confirmed that anaemia is independently associated with adverse outcomes in patients with congestive heart failure, chronic kidney disease, acute coronary syndrome, and sepsis. The prevalence increases with age, approaching 50% in patients who are chronically ill living in nursing homes, and many studies have shown that anaemia is associated with increased morbidity and mortality in patients older than 70 years. Alexander of the study o

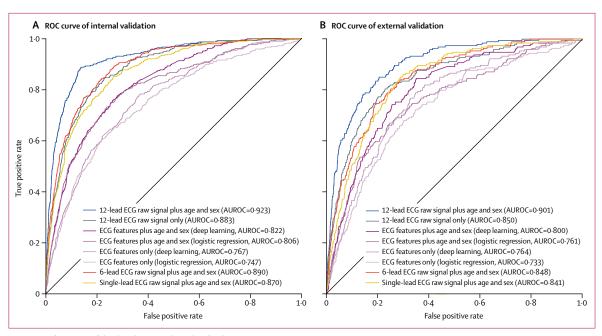
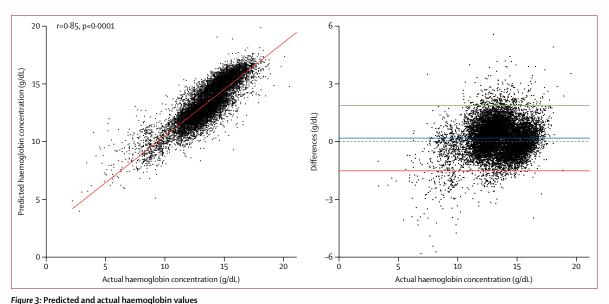


Figure 2: Performance of the deep learning algorithm for detecting anaemia ROC=receiver operating characteristics. AUROC=area under the receiver operating characteristic curve. ECG=electrocardiogram.



The red line is the linear regression line. The blue line is the mean difference, and the red and green lines are the 1-96 SD value from the mean difference.

anaemia is also an important global health problem that affects about 500 million women of reproductive age.²⁵ Delayed detection of anaemia is associated with multi-organ complications. Additionally, the patient could have an acute life-threatening condition, such as gastrointestinal bleeding.³ Most cases of anaemia are easily treated, but the challenge is in screening and diagnosis at an early stage. The ability to non-invasively screen for anaemia using ECG data would help to detect

and monitor the patients at higher risk. As ECGs can be captured by various wearable devices such as patches and watches, the diagnostic ability is especially helpful in low-income countries for screening anaemia without the medical resources required for blood tests.

Using a retrospective dataset of more than $68\,000$ ECGs, the DLA had a high AUROC of 0.923 for internal validation and 0.901 for external validation for identifying anaemia. The model showed good performance with

data from another hospital that was not used for algorithm development and had different patient characteristics and data format. At a high-sensitivity (90%) operating point of development data, the DLA performed well as a potential screening tool to rule out anaemia, with a negative predictive value greater than 99%. The model's performance was better than other common screening tests, such as mammography for breast cancer (AUROC 0·78, positive predictive value 3–12%) and faecal occult blood testing for detecting colorectal neoplasia (AUROC 0·71, overall sensitivity 29%).^{26,27}

The most important aspect of deep learning is its ability to extract features and make an algorithm from various types of data, such as images, 2D data, and waveforms. Here, we used raw ECG data (2D numerical data,

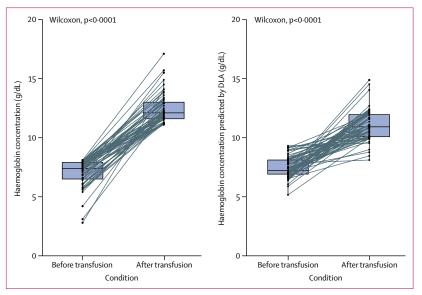


Figure 4: Changes in haemoglobin values and predicted haemoglobin values after transfusion DLA=deep learning algorithm.

12×4000). In previous studies, Attia and colleagues and our study group developed a DLA for screening heart failure, arrhythmia, valvular heart disease, and left ventricular hypertrophy. However, deep learning is often criticised for the unreliability of its outcomes because of the low transparency of the process. Because of this, we used a sensitivity map to visualise the region of the ECG that was used for decision making by the DLA.

The map shows that the DLA focused especially on the QRS complex to decide the presence of anaemia. The DLA also focused on the T-wave for deciding the presence of anaemia. A previous study showed that several electrocardiographic changes had been identified in patients with severe anaemia, including a small QRS amplitude, prolonged QT interval, T-wave inversion, and rightward T-wave axis deviation. 10 In this study, patients with severe anaemia also had a small QRS amplitude, rightward T-wave axis deviation, and longer corrected QT interval, as shown in tables 1, 2. In one case, a patient with anaemia showed a changed ECG after transfusion (appendix p 8). After transfusing a five-pack of red blood cells, the QRS amplitude increased, and the QT interval shortened. When the patient had a status of severe anaemia, bradycardia was present; after transfusion, tachycardia occurred. The predictive values of the DLA decreased from 0.361 to 0.014 (appendix p 8). Most predicted haemoglobin concentrations of patients who were given a transfusion in hospital B increased after transfusion, as shown in figure 4. The reduced myocardial oxygen source might be responsible for the pathophysiological connection between anaemia, with small QRS amplitudes and prolonged QT intervals. In some cases, the DLA used a P wave for its decision. In a previous study, a minor atrioventricular conduction disturbance was associated with anaemia, and patients with severe anaemia had significantly longer PR intervals.11

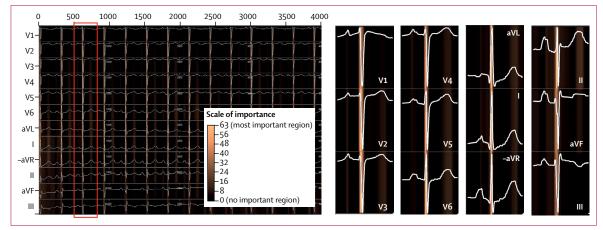


Figure 5: Sensitivity map of a deep learning algorithm for detecting anaemia

The sensitivity map shows the region in which the convolutional neural network algorithm focused attention for deciding the presence of anaemia. The most important region is in orange and the least important region is in black. As the number of filters of the first convolutional layer was 63, the sensitivity map described the region of importance for deciding the presence of anaemia using a 63-grade scale. We visualised grade 0 as black and 63 as orange. V=precordial lead.

Patients with moderate anaemia in this study had increased QRS amplitudes, which could be because those patients had lower body-mass indexes.²⁸

We developed a DLA based on the CNN. We tried the recurrent neural network (RNN) algorithm for this task, but the performance was poor (data not shown). The AUROC for the RNN algorithm was 0.6-0.7. The RNN, such as long short-term memory, was used in some arrhythmia detection tasks in which changing the trend by time was important. In the task in which the shape of each beat in the ECG was important, such as myocardial infarction, heart failure, left ventricular hypertrophy, valvular heart disease, and hyperkalaemia, many researchers use only the CNN. We suggest that CNN is more appropriate than RNN in tasks that change over time, but the shape of the ECG was important. However, as a limitation of the current methodology in deep learning, we could not determine the exact reason why the CNN outperformed RNN in our analyses.

Our study has several limitations to resolve in future work. First, this study was retrospective using conventional 12-lead ECGs. A prospective study is warranted to establish the use of the DLA in patients with the most to gain from a new feasible and non-invasive screening method. A study for confirming the accuracy of the DLA when applied to data from various wearable or portable ECG devices is warranted to apply the DLA to those devices. A study is also needed to confirm the performance of portable ECG devices at home and in daily living. Second, although the performance of the DLA was excellent, the positive predictive value was only 15% at the point of high sensitivity. The positive predictive value and negative predictive value are directly related to the prevalence of the disease in the population. In our study population, the prevalence of severe anaemia was only 0.1% and that of moderate to severe anaemia was only 3.6%. Because ECGs are not usually done to detect anaemia, the prevalence was significantly lower than in the general population in our retrospective study population. We are planning a prospective study of DLA-ECG-anaemia in patients at high risk for anaemia, so the positive predictive value might increase. Furthermore, additional experiments are required to enhance the performance of the DLA using several deep learning methods, such as generative adversarial networks. Third, we need to explore the decision-making process of the DLA further. For example, additional experiments are required to understand the deep learning process better, and thereby understand which exact characteristics of the P wave, QRS complex, and T wave influenced the algorithm's decision. Explainable artificial intelligence (ie, methods for which the results and decision-making processes can be understood by human experts) has been studied and reported on in the past few years, thus the so-called black box limitation (in which humans cannot understand the exact process and reasoning of the artificial intelligence) could be solved in the near future.²⁹ This subject will be our next area of study.

To conclude, in this study a DLA showed high performance for screening anaemia using ECGs. The application of artificial intelligence to the ECG could enable non-invasive screening for anaemia in clinical practice. A prospective study is needed to confirm the DLA using wearable devices in daily living.

Contributors

J-MK contributed to the study idea and design, data collection, prepared and verified the clinical coding, undertook the data analysis, developed the algorithm, wrote the first draft, and contributed to the subsequent drafts. YC and SC did the data analysis and contributed to subsequent drafts. K-HK, SDB, and SJ prepared and verified the clinical coding and did the data analysis. JP and B-HO contributed to the data collection and revised the manuscript. K-HJ is the principal investigator and contributed to the study idea and design, interpreted the data, did the data analysis, and contributed to subsequent drafts. All authors have read and approved the final version of the report.

Declaration of interests

YC and SC are employees of Bodyfriend as researchers of the medical research and development centre. J-MK and JP are co-founders and stakeholders in Medical AI, a medical artificial intelligence company. All other authors declare no competing interests.

Data sharing

Access to datasets from the Sejong General Hospital (hospital A) and Mediplex Sejong Hospital (hospital B), used with permission for this study, should be requested directly from these institutions via their data access request forms. Subject to the institutional review boards' ethical approval, the corresponding author agrees to share de-identified individual participant data, the study protocol, and the statistical analysis plan with academic researchers following completion of a data use agreement. Proposals should be directed to imcardio@gmail.com. The coding used to train the artificial intelligence model are dependent on annotation, infrastructure, and hardware, so cannot be released. However, all experimental and implementation details that can be shared are described in detail in this Article and its appendix. Several major components of our work are available in the TensorFlow open source repository. The corresponding author agrees to share the artificial intelligence algorithm developed from this study with academic researchers following completion of a data use agreement. Proposals should be directed to kwonjm@sejongh.co.kr.

For **TensorFlow open source repository** see https://www. tensorflow.org/

Acknowledgments

This report shows results of a study on the "High Performance Computing Support" Project, supported by the 'Ministry of Science and ICT and National IT Industry Promotion Agency of South Korea.

References

- Mason J, Martorell R, Saldanha L, Shrimpton R. Reduction of anaemia. Lancet Glob Health 2013; 1: e4–6.
- 2 Kassebaum NJ. The global burden of anemia. Hematol Oncol Clin North Am 2016; 30: 247–308.
- Baron BJ, Scalea TM. Acute blood loss. Emerg Med Clin North Am 1996; 14: 35–55.
- 4 Merchant AA, Roy CN. Not so benign haematology: anaemia of the elderly. Br J Haematol 2012; 156: 173–85.
- 5 Tyburski EA, Gillespie SE, Stoy WA, et al. Disposable platform provides visual and color-based point-of-care anemia self-testing. J Clin Invest 2014; 124: 4387–94.
- 6 Kaiafa G, Kanellos I, Savopoulos C, Kakaletsis N, Giannakoulas G, Hatzitolios AI. Is anemia a new cardiovascular risk factor? Int J Cardiol 2015; 186: 117–24.
- 7 Cattadori G, Agostoni P, Corrà U, et al. Heart failure and anemia: Effects on prognostic variables. Eur J Intern Med 2017; 37: 56–63.
- Stucchi M, Cantoni S, Piccinelli E, Savonitto S, Morici N. Anemia and acute coronary syndrome: current perspectives. Vasc Health Risk Manag 2018; 14: 109–18.

- Coppler PJ, Elmer J, Calderon L, et al. Validation of the Pittsburgh Cardiac Arrest Category illness severity score. *Resuscitation* 2015; 89: 86–92
- 10 Avni Findikli H, Sahin Tutak A. The effects of severe iron deficiency anemia on T-wave intervals. Arch Clin Biomed Res 2018; 2: 233–38.
- 11 Gv S, Pk S, Herur A, et al. Correlation between haemoglobin level and electrocardiographic (ECG) findings in anaemia: a crosssectional study. J Clin Diagn Res 2014; 8: BC04–06.
- 12 Francis GS, Tang WHW, Walsh RA. Pathophysiology of heart failure. In: Fuster V, Walsh RA, Harrington RA, eds. Hurst's the Heart, 13th edn. New York, NY: McGraw-Hill, 2011: 482–83.
- 13 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; 521: 436–44.
- 14 Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet* 2019; 394: 861–67.
- 15 Attia ZI, Kapa S, Lopez-Jimenez F, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. Nat Med 2019; 25: 70–74.
- 16 Kwon J-M, Jeon K-H, Kim HM, et al. Comparing the performance of artificial intelligence and conventional diagnosis criteria for detecting left ventricular hypertrophy using electrocardiography. Europace 2019; 22: 412–19.
- 17 Kwon JM, Kim K-H, Akkus Z, Jeon K-H, Park J, Oh B-H. Artificial intelligence for detecting mitral regurgitation using electrocardiography. J Electrocardiol 2020; 59: 151–57.
- 18 Halcox JPJ, Wareham K, Cardew A, et al. Assessment of remote heart rhythm sampling using the AliveCor heart monitor to screen for atrial fibrillation: the REHEARSE-AF study. Circulation 2017; 136: 1784–94.

- 19 Carson JL, Carless PA, Hébert PC. Outcomes using lower vs higher hemoglobin thresholds for red blood cell transfusion. *JAMA* 2013; 309: 83–84.
- 20 Park S, Kwak N. Analysis on the dropout effect in convolutional neural networks. ACCV 2017: 189–204.
- 21 Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015; 10: e0118432.
- 22 Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Proc IEEE Int Conf Comput Vis 2017; 1: 618–26.
- 23 Bross MH, Soch K, Smith-Knuppel T. Anemia in older persons. Am Fam Physician 2010; 82: 480–87.
- 24 Culleton BF, Manns BJ, Zhang J, Tonelli M, Klarenbach S, Hemmelgarn BR. Impact of anemia on hospitalization and mortality in older adults. *Blood* 2006; 107: 3841–46.
- Young MF. Maternal anaemia and risk of mortality: a call for action. Lancet Glob Health 2018; 6: e479–80.
- 26 Pisano ED, Gatsonis C, Hendrick E, et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. N Engl J Med 2005; 353: 1773–83.
- 27 Haug U, Kuntz KM, Knudsen AB, Hundt S, Brenner H. Sensitivity of immunochemical faecal occult blood testing for detecting leftvs right-sided colorectal neoplasia. Br J Cancer 2011; 104: 1779–85.
- 28 Alpert MA, Terry BE, Cohen MV, Fan TM, Painter JA, Massey CV. The electrocardiogram in morbid obesity. Am J Cardiol 2000; 85: 908–10, A10.
- 29 Mittelstadt B, Russell C, Wachter S. Explaining explanations in AI. Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency 2019; 1: 279–88.