EMSNet: A neural network model with a self-attention mechanism for prehospital prediction of care needs.

Joo Jeong[1], Yu Jin Kim[1], Dae Kon Kim[1], Tackeun Kim [2], Joonghee Kim[1*]

[1]Department of Emergency Medicine, [2]Department of Neurosurgery, Seoul National University Bundang Hospital, 166 Gumi-ro, Bundang-gu, Seongnam-si, Gyeonggi-do, 463-707, Republic of Korea.

Correspondence:
Joonghee Kim,
Department of Emergency Medicine
Seoul National University Bundang Hospital
166 Gumi-ro, Bundang-gu, Seongnam-si, Gyeonggi-do, 463-707, Republic of Korea.
E-mail: joonghee@snubh.org

Abstract

Background: An artificial intelligence (AI) system capable of predicting patient needs in the prehospital phase would be instrumental. We sought to develop a neural network (NN) model capable of predicting various care needs at initial contact by emergency medical service (EMS) using multimodal input data.
Methods: We used EMS records of a single emergency department (ED). We implemented two attention-based NN model (I and P) differing only by how they use contextual information. The models predict multiple events, including hospital admission, endotracheal intubation, mechanical ventilation, vasopressor infusion, cardiac catheterization, surgery, intensive care unit (ICU) admission, and cardiac arrest. The input features include both unstructured data (chief complaints, injury summary, past medical history, history of present illness) and structured data (age, sex, pupil status and initial vital signs, level of consciousness, and O2 saturation on pulse oximetry). We applied multi-task learning for training. We evaluated the relative performance of the models compared with a human expert, an emergency physician with 10-year experience as an EMS medical director.
Results: The study population included 42,073 cases. The receiver operating characteristics (ROC) area under the curve (AUC) values of the models I and P ranged from 0.793 to 0.929 and 0.812 to 0.934, respectively. The precision-recall (PR) AUC values ranged from 0.149 to 0.673 and 0.156 to 0.683, respectively. With decision thresholds set to achieve equivalent recall levels, our AI models achieved precision levels not significantly different from those of a human expert except in prediction of mechanical ventilation and ICU admission, where the models achieved superior performance (p=0.030 [model I] and p=0.015 [model P], respectively).
Conclusions: AI models using multimodal input data can predict medical resource requirements at initial contact by EMS with high accuracies.

Keywords
Emergency medical service, Triage, Machine learning, Clinical decision support tool, Prediction, Diagnosis

1. Introduction

Accurate prediction of patients' needs is critical in prehospital care because the type of hospitals and subsequent cares are dependent on it[1,2]. Erroneous ones may lead to inefficient care delivery after transport and, ultimately, poor outcomes. However, the task is not easy. It requires an accurate assessment of both the current condition and the possible future events of a patient, which needs significant knowledge and experience in medicine. Direct medical control by emergency medical services (EMS) directors has been providing essential clues in challenging cases[3–6]. However, it means an increased workload for the directors, and maintaining such support 24/7 is sometimes impossible. Therefore, artificial intelligence (AI) system capable of doing the task with similar performance as a human expert will be a useful resource.

Achieving human expert-level performance requires being able to process unstructured natural language data efficiently as well as structured tabular data. Also, the system should be able to focus on relevant information because natural language data have diverse information.

In this study, we present our self-attention based AI system, EMSNet. It encodes natural language data with contextual information and applies a self-attention mechanism to focus on relevant information. Using the encoded representation and tabular form data, it jointly predicts various hospital resource requirements of a patient. We trained the system using multi-task learning (MTL) methods, and the system achieved human expert-level performance in our experiment.

2. Materials and methods

2.1. Study setting and population

The study is a single-center observational study utilizing EMS records of the patients who visited the emergency department (ED) using public EMS from 2011 to 2015. The study facility is a tertiary academic hospital located in South Korea with an annual ED visits greater than 80,000 patients a year. We excluded out-of-hospital cardiac arrest (OHCA), dead on arrival (DOA), and transferred-out cases. Recurrent visits were treated as independent cases. The institutional review boards of the study site approved the study and provided a waiver of informed consent.

2.2. AI tasks

We applied multi-task learning with one main task and five auxiliary tasks[7]. The main task is a multiple binary prediction problem with its targets include hospital admission, endotracheal intubation, mechanical ventilation, vasopressor infusion, cardiac catheterization, surgery, ICU admission, and cardiac arrest within 24 hours of ED arrival. The auxiliary tasks include the prediction of primary ED diagnosis and ED disposition using the final output of the shared portion of the network (auxiliary task group 1) and performing the main task and the two auxiliary tasks using the intermediate output of the shared network (auxiliary task group 2). The detailed description of the tasks is in supplementary table 1.

2.3. Datasets

We used EMS records of the study facility from 2011 to 2015. We used features obtainable at the initial encounter by EMS to make the AI system to be able to predict the resource needs and thus the destination of transport as soon as possible. The features were age, sex, chief complaints (CC), injury summary (if injury-related), past medical (and surgical) history (PMH), history of present illness (HPI), pupil status (size and reflex), systolic blood pressure (SBP, mmHg), diastolic blood pressure (DBP, mmHg), pulse rate (PR, beats per minute), respiratory rate (RR, breaths per minute) and body temperature (BT, measured in Celsius), level of consciousness (AVPU: Alert, Verbal, Pain and Unresponsive), initial O_2 saturation (SpO_2 on pulse oximetry, %). Free-text data (CC, injury summary, PMH, HPI) were cleaned, lower-cased (for alphabet words), and space-corrected. The target variables of the primary and auxiliary tasks were extracted from the electronic health record (EHR) database.

The dataset was randomly split into training, validation, and test sets with the ratio of 6:2:2. The training dataset was used to develop preprocessing pipelines and to train models. The validation dataset was used to evaluate candidate models and their hyperparameters. The test dataset was used to measure the performance of the final models.

2.4.1. AI system: preprocessing pipelines

The preprocessing pipelines for the multimodal input data are developed using the training dataset and include the following procedures: 1) The NLP pipeline tokenizes and index natural language data using a Korean natural language processing (NLP) tool, soynlp[8]; 2) The pipeline then embeds the tokens using FastText algorithm (implemented in Gensim library) and inhouse-corpus based mainly on Korean Wikipedia and the HPI part of the training dataset[9]. 3) The non-NLP pipeline does the other common preprocessing procedures where categorical features are one-hot encoded, and numerical features are standardized by removing their means and scaling to their unit variances. Missing values are imputed by the means or the modes as appropriate with adding missing indicators to the datasets.

2.4.2. AI system: NN architecture

We assumed a typical healthcare provider would read a medical note in the following sequence, which our EMSNet architecture tries to mimic: First, the reader will briefly look at the contextual information, such as chief complaints, demographics, and underlying conditions. Then the reader will read the whole HPI using the contextual information. Lastly, the reader will interpret various measurements, such as vital signs, exam findings, and test results. If the reader wants to predict a specific health outcome event, he or she can go back to the note and reread the data, focusing on specific parts of the text relevant to the outcome event.

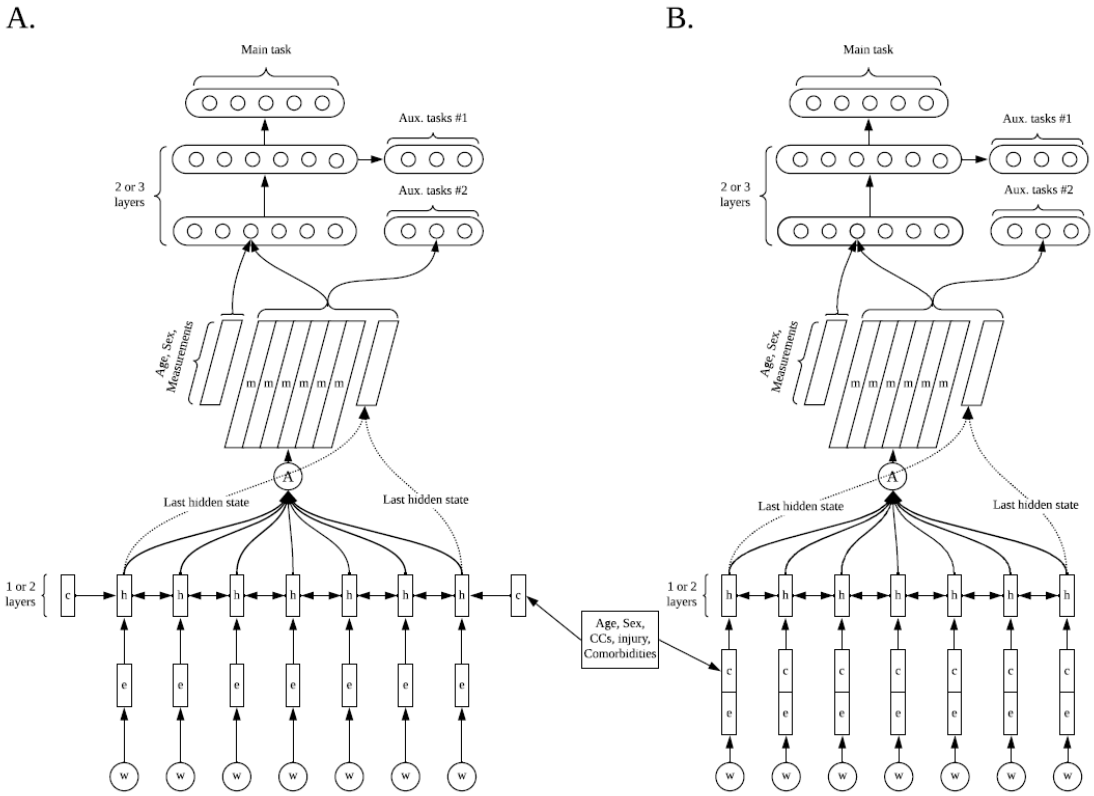


Fig. 1. Network schematics of EMSNet. The model I (A) and P (B) are differed by where the contextual information is incorporated into the computational graph; CC, chief complaint

Figure 1 visualizes the computational graph of EMSNet, which follows a similar process. Briefly, the demographic information (age and sex) and other contextual information (CC, injury summary, and PMH) are concatenated and then transformed by two consecutive fully-connected (FC) layers to output a latent contextual vector $\boldsymbol{c}$. This vector is fed into a l-layer bidirectional gated recurrent unit (GRU) network with a self-attention mechanism where HPI is fused with the contextual information and turned into a sentence embedding. In this process, the contextual information is used in two different ways, either by being overwritten on the initial hidden states of GRUs (model I) or being concatenated with each word embedding vectors of HPI (model P). Then the d_h-dimensional hidden state vectors hs of the GRUs are concatenated at each timestep forming an output matrix H with n-by-$2ld_h$ shape (2 for bidirectional), which is then processed by an attention mechanism proposed by Lin et al.[10] where attention weights $\boldsymbol{a}$ for H is derived from the same H using the following formula:

$$\boldsymbol{a} = softmax(\boldsymbol{w_{s2}}\tanh\,(W_{s1}H^T))$$

Here W_{s1} is a weight matrix with a shape of d_a-by-$2ld_h$ and $\boldsymbol{w_{s2}}$ is a vector of size d_a where d_a is a hyperparameter we can optimize. With $\boldsymbol{a}$, the hidden state matrix H is summed up to obtain a representational vector $\boldsymbol{m}$ of HPI. To obtain multiple representations from a sentence, we extend $\boldsymbol{w}_{s2}$ into a r-by-d_a matrix W_{s2} resulting in the attention weight vector $\boldsymbol{a}$ becoming an attention matrix A. Finally, we obtain a sentence embedding matrix of HPI, M by multiplying A and H, which is then flattened and concatenated with the final hidden state vectors of the GRU network and the standardized measurements (vital signs, consciousness,

pupillary status, and SpO_2) and fed into a sequence of FC layers. Also, there are six task-specific networks with two FC layers, one for the main task and five for the auxiliary tasks.

2.4.3. AI system: loss function

The loss function has two terms. The first one is the weighted sum of cross-entropy losses from the task-specific networks with a weight distribution of 0.5 for the main task and 0.1 for each of the auxiliary tasks. The error signals from auxiliary task group 1 were intended to be used to improve the generalization of the whole network, while those from auxiliary task group 2 were explicitly intended to improve the generalizability of the GRU layers. The other term is a penalizing term P introduced by Lin et al[10].

$$P = \|AA^T - I\|_F^{\ 2}$$

where A is the attention matrix whose rows are attention vectors $\boldsymbol{a}$ as introduced earlier, I is an identity matrix and $\|\cdot\|_F$ stands for the Frobenius norm of a matrix. This term encourages the diversity of the attention vectors $\boldsymbol{a}$ and is multiplied by a hyperparameter we can set arbitrarily.

2.5. Training

The models are trained with Adam optimizer with an early stopping rule requiring five consecutive failures to reduce the minimum loss of the main task. We applied multiplicative learning rate decay after 5th, 10th, 15th and 30th epoch with its decay rate parameterized for optimization. Supplementary Table 2 shows all the hyperparameters we used. They were optimized using tree-structured Parzen estimation with over 500 trials for each model[11].

2.6. Performance evaluation

We assessed the performance of our AI systems measuring receiver operating characteristic (ROC) and precision-recall (PR) area under the curve (AUC) values. We used bootstrap resampling (N=2000) to calculate the 95% confidence intervals (CI) of the AUC values and to test the statistical significance of their differences. We set up a human expert vs. AI competition to evaluate the performance of our models. For this competition, we randomly sampled up to 10 cases without replacement from each outcome combination stratum (N=313), and then we added 200 additional cases sampled without stratification and replacement from the rest finalizing our final competition dataset (N total=513). The stratified sampling procedure was to increase the proportion of positive outcome cases, which will increase the statistical power of later comparison tests. A board-certified emergency medicine (EM) physician with 10-year experience as an EMS director predicted outcomes using the competition dataset. Using the results, we calculated recall (sensitivity) levels of the human expert and set the threshold levels of our models to achieve the same recall levels as the human expert. Lastly, we calculated and compared precision (positive predictive value, PPV), negative predictive value (NPV), and specificity levels of the models and the human expert. Their 95% confidence intervals and the significance of difference was assessed using bootstrap resampling.

2.7. Visualization and quality assessment of attention mapping

We visualize where our models are focusing on by drawing a heatmap over the tokens of the HPI sentences. The values of the heatmap are obtained by summing over all the attention vector $\boldsymbol{a}$ and rescaling the vector using min-max normalization (ranging 0 to 1).

A separate reviewer (a board-certified EM physician with two years of EMS director experience) rated the clinical relevance of the attention patterns in one hundred random samples from the test dataset using a 5-point Likert scale (Perfect, Good, Fair, Poor, Random). The reviewer was instructed to determine the quality based on general clinical relevance without being instructed for what purpose the models are used.

2.8. Statistical analysis

Categorical variables are reported using frequencies and proportions. Continuous variables are reported using the median and interquartile range (IQR). T-test, Wilcoxon's rank-sum test, chi-square test, or Fisher's exact test are performed as appropriate for comparison between groups.

P-values < 0.05 were considered significant. Neural network models were developed and tested using PyTorch package version 1.4 running on Python version 3.7[12]. Statistical analyses were performed on R-packages version 3.5.1 (R Foundation for Statistical Computing, Vienna, Austria).

3. Results

45,396 ED visits using the national EMS were identified. After the exclusion of OHCA, DOA, and transfer-out cases, 42,073 cases were included as the study population (Table 1). The median age of the population was 58.0 (43.0-73.0), and female cases were 21,023 (50.0%). The main outcome events including hospital admission, endotracheal intubation, mechanical ventilation, pressor infusion, surgery, cardiac catheterization, ICU admission and cardiac arrest occurred in 10,689 (25.4%), 915 (2.2%), 808 (1.9%), 1402 (3.3%), 1472 (3.5%), 783 (1.9%), 2226 (5.3%) and 310 (0.7%) cases, respectively. The number of cases in train, validation, and test dataset was 25,242, 8,414, and 8,414, respectively with no significant difference among the groups.

		Total (N=42073)	Train (N=25245)	Validation (N=8414)	Test (N=8414)	p
Demographics	Age	58.0 (43.0-73.0)	58.0 (43.0-73.0)	58.0 (43.0-73.0)	58.0 (43.0-73.0)	0.638
	Sex					0.745
	- Female	21023 (50.0%)	12635 (50.0%)	4215 (50.1%)	4173 (49.6%)	
	- Male	21050 (50.0%)	12610 (50.0%)	4199 (49.9%)	4241 (50.4%)	
Initial measurements	SBP	130.0 (116.0-150.0)	130.0 (116.0-150.0)	130.0 (117.0-150.0)	130.0 (116.0-149.0)	0.620
	DBP	80.0 (70.0-90.0)	80.0 (70.0-90.0)	80.0 (70.0-90.0)	80.0 (70.0-90.0)	0.050
	Pulse rate	82.0 (72.0-95.0)	82.0 (72.0-95.0)	82.0 (72.0-95.5)	82.0 (72.0-96.0)	0.475
	Respiratory rate	18.0 (16.0-20.0)	18.0 (16.0-20.0)	18.0 (16.0-20.0)	18.0 (16.0-20.0)	0.236
	Body temperature	36.5 (36.2-36.9)	36.5 (36.2-36.9)	36.5 (36.2-36.9)	36.5 (36.2-36.9)	0.692
	SpO2	98.0 (96.0-99.0)	98.0 (96.0-99.0)	98.0 (96.0-99.0)	98.0 (96.0-99.0)	0.407
	BST	132.0 (104.0-182.0)	132.0 (104.0-181.0)	133.0 (104.5-181.5)	134.0 (105.0-186.0)	0.606
	Pupil size (left)	3.0 (3.0-3.0)	3.0 (3.0-3.0)	3.0 (3.0-3.0)	3.0 (3.0-3.0)	0.295
	Pupil size (right)	3.0 (3.0-3.0)	3.0 (3.0-3.0)	3.0 (3.0-3.0)	3.0 (3.0-3.0)	0.203
	Pupil status (left)					0.112
	- Normal	40427 (96.1%)	24303 (96.3%)	8053 (95.7%)	8071 (95.9%)	
	- Dilated	190 (0.5%)	115 (0.5%)	33 (0.4%)	42 (0.5%)	
	- Constricted	270 (0.6%)	156 (0.6%)	65 (0.8%)	49 (0.6%)	
	- Missing	1186 (2.8%)	671 (2.7%)	263 (3.1%)	252 (3.0%)	
	Pupil status (right)					0.044
	- Normal	40381 (96.0%)	24290 (96.2%)	8037 (95.5%)	8054 (95.7%)	

	- Dilated	189 (0.4%)	111 (0.4%)	35 (0.4%)	43 (0.5%)	
	- Constricted	272 (0.6%)	153 (0.6%)	67 (0.8%)	52 (0.6%)	
	- Missing	1231 (2.9%)	691 (2.7%)	275 (3.3%)	265 (3.1%)	
	Light reflex (left)					0.807
	- Reactive	40439 (96.1%)	24285 (96.2%)	8075 (96.0%)	8079 (96.0%)	
	- Unreactive	297 (0.7%)	172 (0.7%)	58 (0.7%)	67 (0.8%)	
	- Unmeasurable	799 (1.9%)	464 (1.8%)	169 (2.0%)	166 (2.0%)	
	- Missing	538 (1.3%)	324 (1.3%)	112 (1.3%)	102 (1.2%)	
	Light reflex (right)					0.382
	- Reactive	40375 (96.0%)	24267 (96.1%)	8049 (95.7%)	8059 (95.8%)	
	- Unreactive	301 (0.7%)	167 (0.7%)	64 (0.8%)	70 (0.8%)	
	- Unmeasurable	789 (1.9%)	453 (1.8%)	169 (2.0%)	167 (2.0%)	
	- Missing	608 (1.4%)	358 (1.4%)	132 (1.6%)	118 (1.4%)	
	Initial Consciousness					0.489
	- Alert	39339 (93.5%)	23593 (93.5%)	7849 (93.3%)	7897 (93.9%)	
	- Verbal	1656 (3.9%)	1010 (4.0%)	337 (4.0%)	309 (3.7%)	
	- Pain	904 (2.1%)	543 (2.2%)	194 (2.3%)	167 (2.0%)	
	- Unresponsive	174 (0.4%)	99 (0.4%)	34 (0.4%)	41 (0.5%)	
	MEWS	1.0 (1.0-2.0)	1.0 (1.0-2.0)	1.0 (1.0-2.0)	1.0 (1.0-2.0)	0.483
Hospital outcomes	Admitted	10689 (25.4%)	6434 (25.5%)	2111 (25.1%)	2144 (25.5%)	0.757
	Endotracheal intubation	915 (2.2%)	538 (2.1%)	175 (2.1%)	202 (2.4%)	0.272
	Mechanical ventilation	808 (1.9%)	476 (1.9%)	151 (1.8%)	181 (2.2%)	0.197
	Pressor infusion	1402 (3.3%)	811 (3.2%)	298 (3.5%)	293 (3.5%)	0.240
	Surgery	1472 (3.5%)	857 (3.4%)	328 (3.9%)	287 (3.4%)	0.083
	Cardiac catheterization	783 (1.9%)	489 (1.9%)	150 (1.8%)	144 (1.7%)	0.348
	ICU admission	2226 (5.3%)	1327 (5.3%)	447 (5.3%)	452 (5.4%)	0.915
	Cardiac arrest in 24 hours	310 (0.7%)	185 (0.7%)	60 (0.7%)	65 (0.8%)	0.897
	ED Disposition					0.509
	- Discharge	31831 (75.7%)	19096 (75.6%)	6388 (75.9%)	6347 (75.4%)	
	- Ward	7839 (18.6%)	4739 (18.8%)	1528 (18.2%)	1572 (18.7%)	
	- ICU	1724 (4.1%)	1027 (4.1%)	351 (4.2%)	346 (4.1%)	
	- OR	463 (1.1%)	255 (1.0%)	107 (1.3%)	101 (1.2%)	
	- Death	216 (0.5%)	128 (0.5%)	40 (0.5%)	48 (0.6%)	

Table 1. Study population

Figure 2 and 3 shows ROC and PR curves of the models, respectively, assessed in the test dataset. The ROC AUC values of the model I and P ranged from 0.793 to 0.929 and 0.812 to 0.934. respectively, both of which were higher to those of modified early warning score (MEWS) in every aspect (all $p < 0.001$, supplementary Table 3). PR AUC values ranged from 0.149 to 0.673 and 0.156 to 0.683, respectively, and were higher to those of MEWS (all $p < 0.001$). The model P generally performed better with significantly higher ROC AUC in the prediction of admission ($p=0.017$), mechanical ventilation ($p=0.028$), surgery ($p=0.011$), and cardiac arrest ($p=0.006$) and with significantly higher PR AUC in the prediction of admission ($p<0.001$), intubation ($p=0.010$), mechanical ventilation ($p=0.005$), and ICU admission ($p=0.010$).

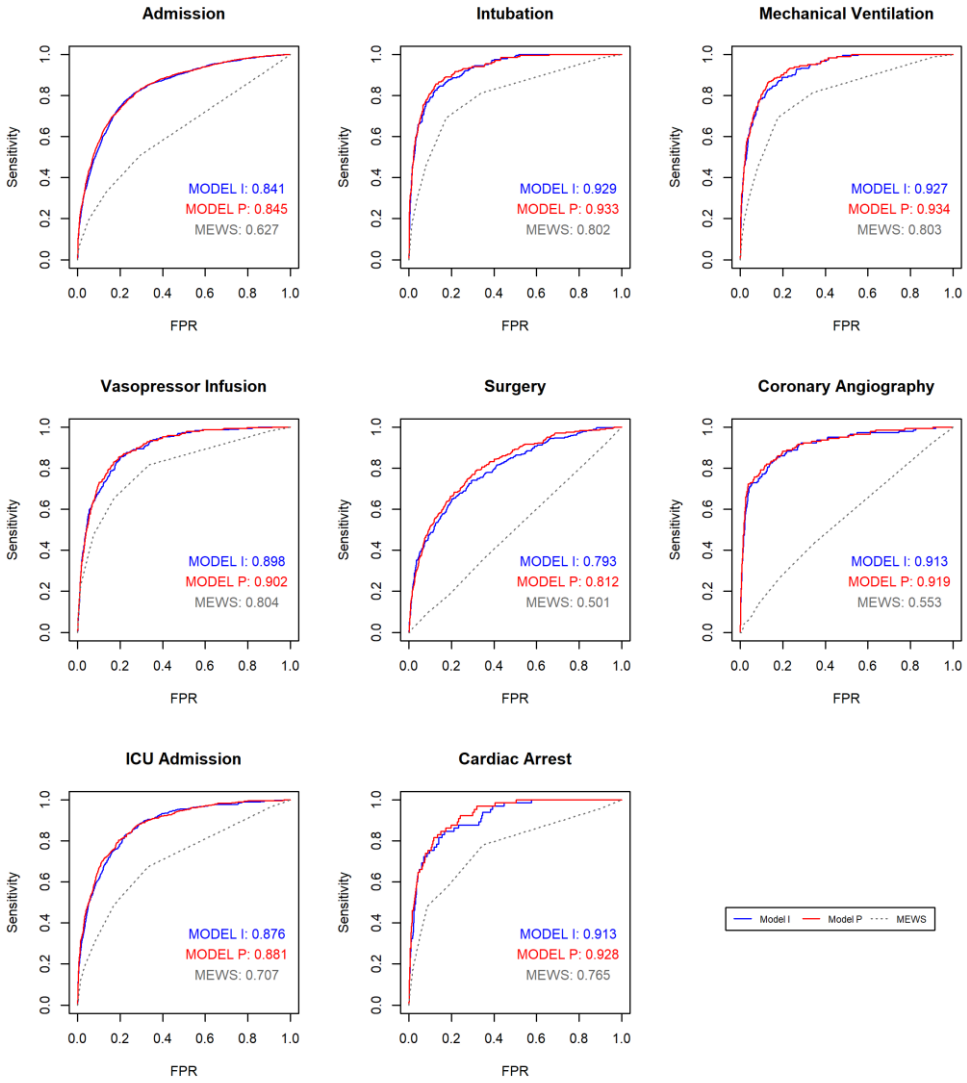


Fig. 2. Receiver operating characteristic (ROC) curves of the model I (blue lines) and P (red lines) plotted against those of modified early warning score (MEWS, gray dashed lines) and their area under the curve (AUC) values; ICU, intensive care unit

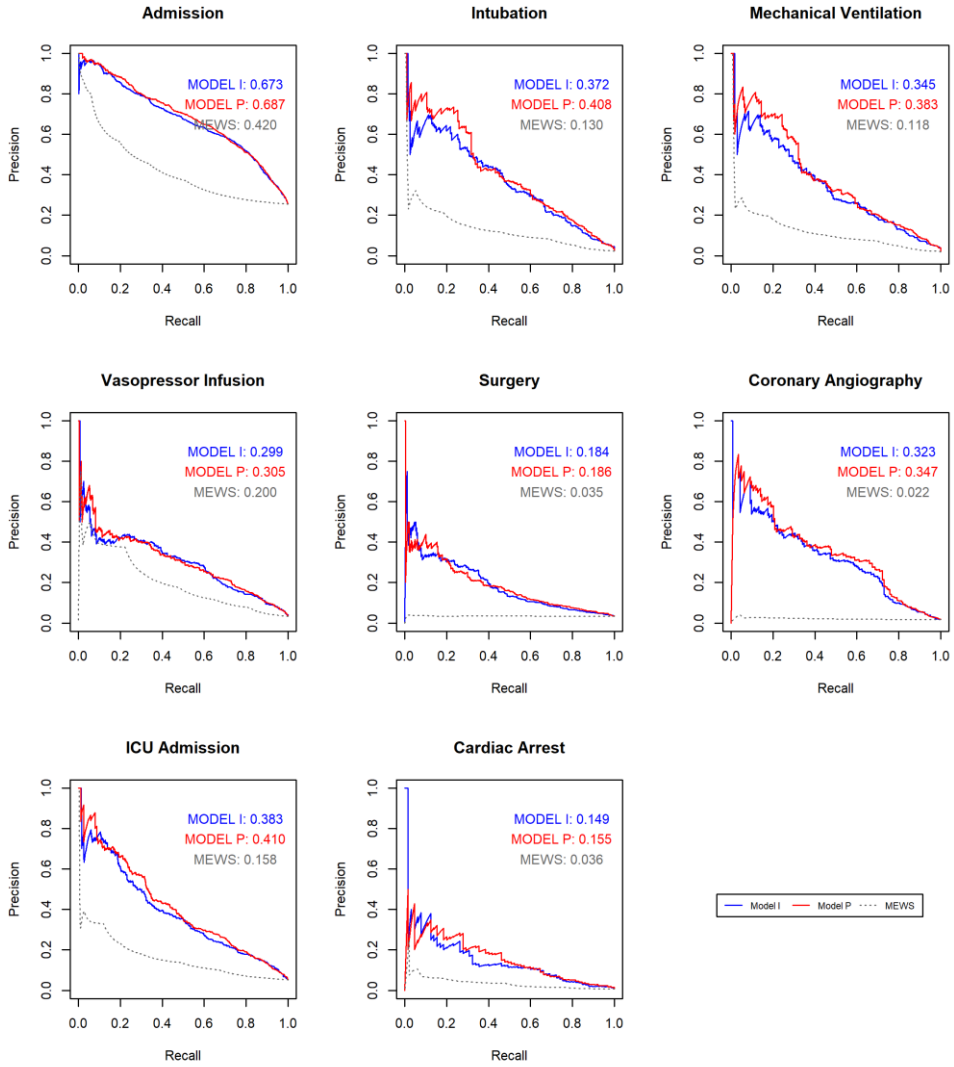


Fig. 3. Precision-recall (PR) curves of the model I (blue lines) and P (red lines) plotted against those of modified early warning score (MEWS, gray dashed lines) and their area under the curve (AUC) values; ICU, intensive care unit

Figure 4 shows the results of a human expert vs. AI competition test. Our AI models achieved precision levels not significantly different from those of a human expert except in prediction of mechanical ventilation and ICU admission, where they achieved superior performance (p=0.030 [model I] and p=0.015 [model P], respectively, supplementary Table 4).

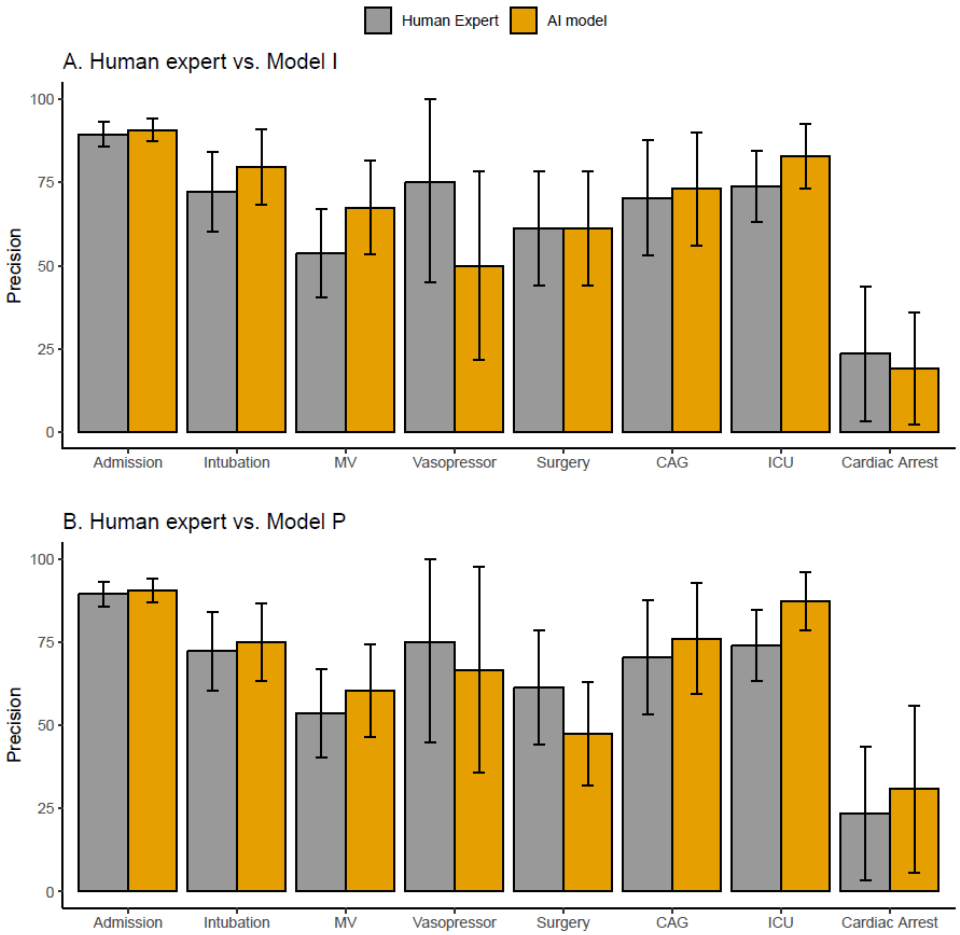


Fig. 4. Precision comparison in the human expert vs. AI competition;
MV, mechanical ventilation; CAG, coronary angiography; ICU, intensive care unit

Another human expert with two years of EMS director experience rated the quality of attention mappings in one hundred random samples from the test dataset (Figure 5). Only 10 percent of the cases were rated poor or worse (poor: 8, random: 2). Supplementary figure 1 shows the representative samples of attention maps chosen from the perfect/good/fair category cases.

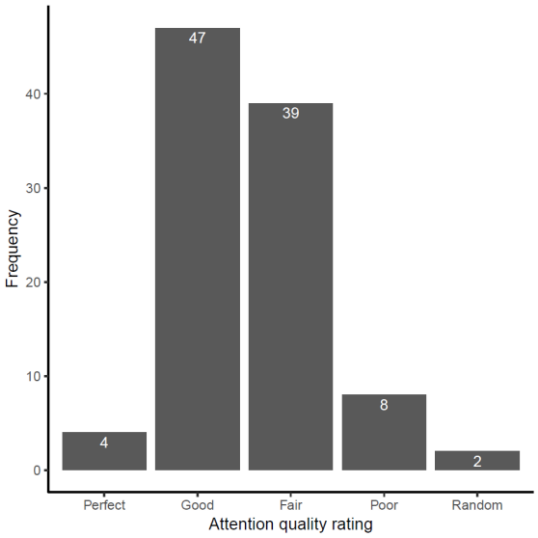


Fig. 5. Quality ratings of attention mappings by a human expert in sample cases (N=100)

4. Discussion

In this study, we designed AI models that can jointly predict various hospital care needs using multimodal data at initial contact by EMS. The self-attention based-models were trained with multi-task learning methods. Our experiment showed that AI models could achieve similar or better performance than a human expert in this domain.

Accurate prediction of patients' needs for hospital resources is critical because the type of hospitals and subsequent cares are dependent on it. [2,13]Direct medical control can improve the quality of the prediction. However, it requires 24/7 access to EMS directors and may lead to increased workload. Several tools have been developed to help the decision process where a single yes or no type event (i.e., mortality or ICU admission) is predicted based on limited types and number of variables[14–16]. In this approach, one can achieve near-maximum performance allowed by the dataset using shallow algorithms (i.e., logistic regression with polynomial and interaction terms or other non-deep learning-based ML methods) if developed in a principled way[17,18].

This approach, however, has obvious limitations. First, in many medical emergencies, a patient commonly has multiple care needs that cannot be predicted by a single output. Also, some of these predictions (i.e., surgery or coronary angiography) require target-specific features often scattered around in natural language data or in other unstructured data forms. In short, we need multiple outputs, each of which focusing on relevant information from both structured and unstructured data. This requirement motivated our adoption of self-attention mechanism. In the self-attention mechanism, the focus of attention is not hard-coded and determined by the input representation and queries. Through training, the mechanism learns to generate multiple attention patterns, each conditioned by each of the queries. The queries are also learnable and were parameterized by the W matrix in our models. How many queries a model needs for optimal performance would be task- and data-specific. The Bayesian hyperparameter optimization procedure suggested models with relatively many queries (eleven for both of the model types).

The size of our training dataset was relatively small. Considering the rare occurrence of the outcome events, it is surprising that our models achieved performance similar to a human expert. One of the possible reasons could

be the use of multi-task learning. Health outcomes are often highly related to one another. One can take advantage of this by using multi-task learning, where a shared intermediate representation is used for multiple tasks. Possible benefits have been suggested[7]: 1) It has implicit data augmentation effect; 2) It helps the models to focus more on relevant features rather than noises; 3) It allows a task to use the features developed by another task; 4) It biases a model to prefer representations that other tasks also prefer which will help the model to be generalized to new tasks; 5) It acts as a regularizer by introducing an inductive bias which reduces the risk of overfitting. In our study, the feature-rich natural language input data could have made the models very prone to overfit. Our extensive application MTL may have provided significant benefits, probably in data augmentation and model regularization.

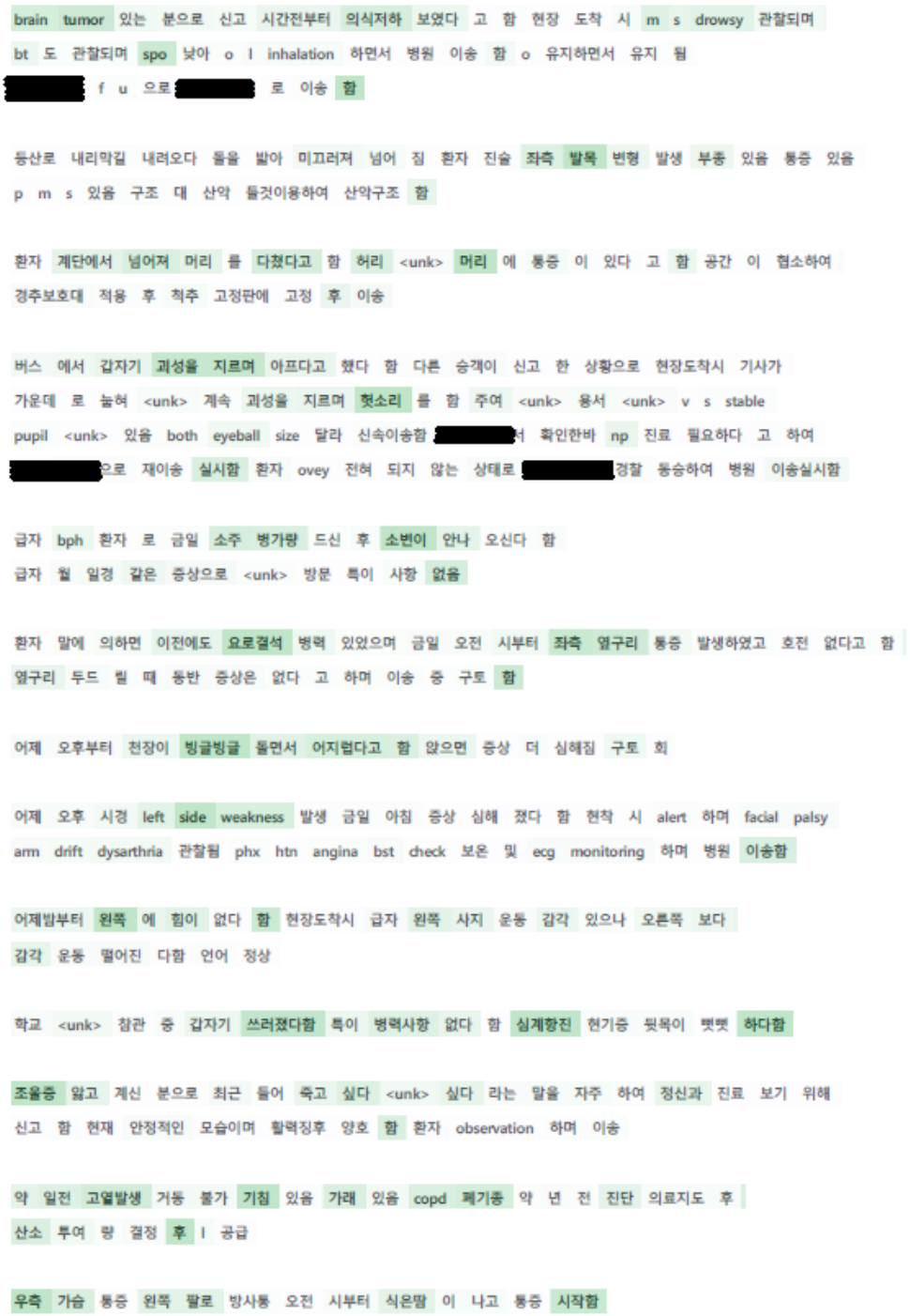
The model P performed better than model I. This suggests the contextual information should be repeatedly provided with each new input vector rather than being used once at the beginning of the unrolling of GRUs. This could be explained by that the contextual information encoded in the initial hidden state will degrade as the unrolling progress and be "forgotten" eventually.

This study has several limitations. Firstly, the models were developed and tested using EMS records of a single hospital. The way of using natural language, as well as the population characteristics and clinical environment, can change by time and space. Therefore, we cannot guarantee the generalizability of our models. Second, only one human expert was compared to the AI models. We chose to use the performance scores of the most experienced EMS director as a comparator rather than averaged scores. Third, the role of direct medical control is much wider rather than some predictions of hospital care needs. Developing a system capable of providing all the expertise of an EMS director would be much more challenging.

Despite these limitations, our study has several strengths. This is the first study developing AI systems capable of jointly predicting multiple outcomes using only prehospital information. The system achieved human expert-level performance and provided interpretable outputs. Also, this is the first attempt to apply modern NLP techniques on EMS records. The models extract distributed sentence representations from unstructured real-world free text data were used for predictions. We also presented some clues on how contextual information can be incorporated into the computational graphs to improve prediction performance. Lastly, we also showed how multiple auxiliary tasks can be utilized in model development.

5. Conclusion

Our models with a self-attention mechanism trained using a multi-task learning method achieved similar (or superior) performance compared to an experienced human expert. Our study shows that AI models can be used to predict various medical resource requirements at initial contact by EMS.



Supplementary Fig. 1. Examples of attention mappings generated by the models

Acknowledgements

This work was supported by SNUBH grant No. 02-2013-060 and SNUBH grant No. 09-2015-001.

Disclosures

The authors declare no conflict of interest.

References

1. Moore L: Measuring quality and effectiveness of prehospital ems. Prehosp Emerg Care. 1999;3(4):325–31.
2. Seymour CW, Kahn JM, Cooke CR, et al.: Prediction of Critical Illness During Out-of-Hospital Emergency Care. Jama. 2010;304(7):747.
3. Role of the State EMS Medical Director. Ann Emerg Med. 2017;70(1):110–2.
4. O'Connor RE, Ali AS, Brady WJ, et al.: Part 9: Acute Coronary Syndromes: 2015 American Heart Association Guidelines Update for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care. Circulation. 2015;132(18 Suppl 2):S483-500.
5. Andersson U, Söderholm HM, Sundström BW, et al.: Clinical reasoning in the emergency medical services: an integrative review. Scand J Trauma Resusc Emerg Medicine. 2019;27(1):76.
6. Moore B, Shah MI, Owusu-Ansah S, et al.: Pediatric Readiness in Emergency Medical Services Systems. Prehosp Emerg Care. 2019;24(2):175–9.
7. Ruder S: An Overview of Multi-Task Learning in Deep Neural Networks. Arxiv.
8. Kim H: soynlp.
9. Bojanowski P, Grave E, Joulin A, et al.: Enriching Word Vectors with Subword Information. Arxiv.
10. Lin Z, Feng M, Santos C dos, et al.: A Structured Self-attentive Sentence Embedding.
11. Bergstra J, Yamins D, Cox D: Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms. doi: 10.25080/majora-8b375195-003 (Epub ahead of print).
12. Paszke A, Gross S, Massa F, et al.: PyTorch: An Imperative Style, High-Performance Deep Learning Library. Arxiv.
13. Kahn JM, Branas CC, Schwab CW, et al.: Regionalization of medical critical care: What can we learn from the trauma experience?*. Crit Care Med. 2008;36(11):3085–8.
14. Kang D-Y, Cho K-J, Kwon O, et al.: Artificial intelligence algorithm to predict the need for critical care in prehospital emergency medical services. Scand J Trauma Resusc Emerg Medicine. 2020;28(1):17.
15. Baxt WG, Jones G, Fortlage D: The trauma triage rule: A new, resource-based approach to the prehospital identification of major trauma victims. Ann Emerg Med. 1990;19(12):1401–6.
16. Lidal IB, Holte HH, Vist GE: Triage systems for prehospital emergency medical services - a systematic review. Scand J Trauma Resusc Emerg Medicine. 2013;21(1):28.
17. Trevor H, Robert T, Friedman J: The elements of statistical learning: data mining, inference, and prediction. New York, NY: Springer.
18. Bishop CM: Pattern recognition and machine learning. 2006.