Genetics in Medicine (2022) 24, 2329–2337

ELSEVIER

Genetics in Medicine
An Official Journal of the ACMG

www.journals.elsevier.com/genetics-in-medicine

ARTICLE

Using deep learning and electronic health records to detect Noonan syndrome in pediatric patients

Zeyu Yang[1,2], Amy Shikany[3], Yizhao Ni[1,2], Ge Zhang[2,4], K. Nicole Weaver[2,3,4], Jing Chen[1,2,*]

[1]*Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati Children's Research Foundation, Cincinnati, OH;* [2]*Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH;* [3]*Heart Institute, Cincinnati Children's Hospital Medical Center, Cincinnati, OH;* [4]*Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH*

ARTICLE INFO

Article history:
Received 2 February 2022
Received in revised form
3 August 2022
Accepted 4 August 2022
Available online 13 September 2022

Keywords:
Deep learning
Diagnostics
Electronic health records
Mendelian disorder
Noonan syndrome

ABSTRACT

Purpose: The variable expressivity and multisystem features of Noonan syndrome (NS) make it difficult for patients to obtain a timely diagnosis. Genetic testing can confirm a diagnosis, but underdiagnosis is prevalent owing to a lack of recognition and referral for testing. Our study investigated the utility of using electronic health records (EHRs) to identify patients at high risk of NS.

Methods: Using diagnosis texts extracted from Cincinnati Children's Hospital's EHR database, we constructed deep learning models from 162 NS cases and 16,200 putative controls. Performance was evaluated on 2 independent test sets, one containing patients with NS who were previously diagnosed and the other containing patients with undiagnosed NS.

Results: Our novel method performed significantly better than the previous method, with the convolutional neural network model achieving the highest area under the precision-recall curve in both test sets (diagnosed: 0.43, undiagnosed: 0.16).

Conclusion: The results suggested the validity of using text-based deep learning methods to analyze EHR and showed the value of this approach as a potential tool to identify patients with features of rare diseases. Given the paucity of medical geneticists, this has the potential to reduce disease underdiagnosis by prioritizing patients who will benefit most from a genetics referral.

Introduction

Noonan syndrome (NS), including Noonan spectrum disorders, is one of the most common syndromic causes of congenital heart defect, second only to Down syndrome.[1] Caused by the pathogenic variants in multiple genes of the RAS/MAPK pathway, NS is a primarily autosomal dominant disorder characterized by short stature, distinctive facial features, and cardiac abnormalities.[2,3] Clinical diagnosis of NS can be difficult because of the

[*]Correspondence and requests for materials should be addressed to Jing Chen, Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, Cincinnati, OH 45229-3026. *E-mail address:* jing.chen2@cchmc.org

doi: https://doi.org/10.1016/j.gim.2022.08.002

heterogeneity in systems affected and variable severity of phenotypes. Genetic testing, such as gene panels and exome sequencing, can identify causal pathogenic variants in 70% to 80% of the patients with NS.[4] However, only a small portion of individuals undergo such testing because of the difficulty in recognition and lack of referral for genetic testing.[5] In Cincinnati Children's Hospital Medical Center's (CCHMC) electronic health record (EHR) database, 198 patients were found to be diagnosed with NS among the 1 million patients (a prevalence of about 1 in 5000 patients) in the database in 2019. A similar incidence rate was noticed in March 2021 in which 1086 NS cases were found among the 6.4 million patients (a prevalence of about 1 in 6000 patients) in PEDSnet (with CCHMC patients excluded), a network formed by 8 pediatric hospitals in the United States.[6] The prevalence of NS in CCHMC and PEDSnet is significantly lower than the estimated prevalence of NS in the US population, which is about 1 in 1000 to 2500 live births.[3,7] This finding raised the concern that there could be a substantial number of undiagnosed patients with NS in CCHMC and similar pediatric institutions. The aim of our study was to develop a computational method to detect patients at high risk for NS on the basis of EHR data.

To better use the information present in the EHR system, we chose to use EHR diagnosis description texts as our training data and adopted a supervised machine-learning approach that focused on NS as the prediction target. In our previous study,[8] we developed a computational method named Genetic Disease Diagnosis based on Phenotypes (GDDP) to match the patient's phenotypes extracted from diagnosis description text in EHR and represented as Human Phenotype Ontology (HPO[9]) terms against the genetic diseases in the OMIM database.[10] Although GDDP took account of the hierarchical structure of HPO terms and their information content and achieved superior performance compared with previous methods,[8] it was designed to facilitate phenotype-guided genetic testing for individual patients rather than prioritizing patients for a specific genetic disease. Because more clinical information is available in EHR data, we hypothesized that better predictive performance could be achieved by directly training machine-learning models using EHR data.

Previous studies have tried using facial imaging approaches to discriminate NS from other genetic diseases.[11-13] However, to our knowledge, no study has reported using a text-based deep learning approach to systemically detect NS. In this study, we focused on deep learning methods, because they are especially efficient at learning from text, while eliminating the need for explicit natural language processing and manual feature selection. Using deep learning methods with EHR diagnosis description text as input, therefore, could significantly simplify the data preprocessing, model training, and result interpretation, all of which reduce the cost of model development and increase the generalizability of the method.

Materials and Methods

To build and evaluate our proposed framework to detect NS from EHR, we compiled a training set consisting of 162 NS cases and 16,200 presumed controls and evaluated the performance of the models based on 2 independent test sets containing diagnosed and undiagnosed NS cases separately. Details of the steps, model performance, and clinical interpretation of results will be described in the following sections. This study protocol was reviewed and approved by the Cincinnati Children's Hospital Institutional Review Board (protocol number 2020-0685).

Data description

Two types of diagnosis data are available in the CCHMC de-identified EHR database (Informatics for Integrating Biology and the Bedside, https://www.i2b2.org/). The first type is diagnosis code, such as the International Classification of Diseases-10th revision.[14] The second type is diagnosis description text, which is based on the Intelligent Medical Objects terminology. The diagnosis description texts were used as the input data for this project, which provided a richer and more accurate description of phenotypes than diagnosis codes.[15] The diagnosis description text found in encounter, problem list, and billing, together with the patient's de-identified identifier and gender were exported from EHR on May 21, 2019.

Identification of patients with NS and data preparation

A total of 198 patients with NS were identified in the 2019 data set through a text search in the diagnosis description. Patients with "Noonan syndrome" or "Noonan's syndrome" (ignoring case distinctions) in the diagnosis description were selected. The discovered patients matched with a recent retrospective study conducted in the institution,[16] confirming that patients with NS identified were accurate. A total of 9 patients with less than 10 words in their diagnosis description were excluded. The remaining 189 patients were labeled as NS cases and were randomly split into 2 groups, with 162 and 27 in each group (a ratio of 6:1). The first group (162 NS cases) was used as cases for training, and the second group (27 NS cases) was placed into our first test set, referred to as test set 1. We labeled the remaining patients who did not have NS diagnosis and possessed at least 10 words in their diagnosis description as the presumed controls. It should be noted that there could be a small number of unrecognized NS cases in the presumed controls. However, our simulations suggested that their existence had minimal effect on the models' training and testing results.

Diagnosis descriptions were sequentially added together to form a single string of descriptions for each patient. The gender of the patient was added to the beginning of each description string. The diagnoses in each string were ordered in the same manner as the original records, with the earlier diagnoses at the beginning and the most recent at the end of the string. A tokenizer was used to convert every word in the description text vocabulary (containing 15,725 unique words) to a unique integer representation. Because 99.6% of strings were shorter than 1000 terms, the first 1000 integers were used as the input features for all patients. All tokenized sequences were right-padded or truncated to be 1000 integers long.

Diagnoses containing the term "Noonan" or associated with International Classification of Diseases-10 codes "Q87.1" and "Q87.19" (indicating clinically diagnosed NS or other genetic diseases associated with short stature), and "Z00.6" (patients in clinical research, exam for clinical research, etc) were removed from all samples before performing training and testing because these terms could be confounded with NS diagnosis.

Training and cross-validation

We performed 6-fold cross-validation to optimize the hyperparameters of the models, using 5 folds (135 cases) for training and 1 fold (27 cases) for validation in each iteration. For each training set, we included 13,500 random control samples so that the case to control ratio was 1:100. The prevalence of NS cases in training was artificially increased to help the models learn the features of NS cases more efficiently, because including more controls in the training set slowed down the training process and did not result in significant improvement in performance. For the validation set, we included 27,000 random control samples so that the case to control ratio was 1:1000. The case to control ratio of 1:1000 in the validation set was selected to approximate the expected prevalence of NS in the general population. Owing to the imbalanced nature of the problem, we used precision, recall, F1 score, and area under precision-recall curve (PRAUC) as our model performance metrics instead of sensitivity, specificity, and area under receiver-operating characteristic curve.[17]

Deep learning models

We tested multiple deep learning techniques, including multilayer perceptron (MLP),[18] convolution neural net (CNN),[19] and recurrent neural net (RNN)[19] (long short-term memory [LSTM],[20] gated-recurrent units [GRU],[21] and their bidirectional variants). MLP is a type of densely connected network. Whereas MLP learns global patterns, CNN can learn local patterns and has been used for text classification and image recognition.[22] RNN keeps an internal memory of information and has been crucial for the field of natural language processing and time series data analysis.[23] LSTM and GRU are more sophisticated variants of the base RNN that use memory gates to mediate the retention of information.

All models started with an embedding layer that converts the tokenized terms into 128D vector representations. GloVe pretrained embedding[24] was also tested but performed poorly and was not used in final models. The hidden layers varied depending on the type of model used. The final layer was a single-neuron dense layer with sigmoid activation. Binary cross-entropy loss was used for all models. Multiple regularization techniques were used to reduce overfitting, including dropout, L1, L2 regularization, and early stopping (implementation details on GitHub). The final model was not retrained using all of the data because we chose to use early stopping.

Performance evaluation using 2 different test sets

The performance of the final model was evaluated based on 2 independent test sets. Test set 1 contained 27 diagnosed NS cases and 27,000 random control samples that were not used in the training, with a case to control ratio of 1:1000. The case to control ratio was selected for the same reason as the validation set.

Test set 2 contained 10 undiagnosed NS cases and 10,000 random controls. We queried the CCHMC EHR database on October 21, 2021, and found 11 patients who were present in our 2019 data set but were diagnosed with NS during 2020 and 2021. These 11 patients confirmed our suspicion that there were patients with undiagnosed NS in our 2019 data. Because 1 of the 11 patients was used in training our final model as a presumed control, he/she was excluded from test set 2.

Explanation and interpretation of the model

To derive an interpretation of the model for NS detection, each unique diagnosis description text was treated like an individual patient and evaluated by the final model. The description texts were ranked by their prediction scores to represent their relative importance for NS detection. To evaluate the importance of word order in each diagnosis text, we also ran the final model on each of the unique diagnosis description text with permuted word orders. Specifically, each diagnosis description was permuted multiple times to derive multiple permuted scores. A one-sided one-sample Wilcox rank-sum test of the permuted scores was then performed against the original score to derive the P value. A P value of $< .05$ indicated the score from the original diagnosis description text significantly higher than that from the permuted diagnosis.

To understand how the diagnostic decision was made for each individual patient by the deep learning diagnostic model, we performed sensitivity analysis for each patient in which we examined the change in prediction score when each diagnosis text was removed from the input. This analysis was similar to LIME[25] but considered diagnosis texts instead of individual words as features. The diagnosis

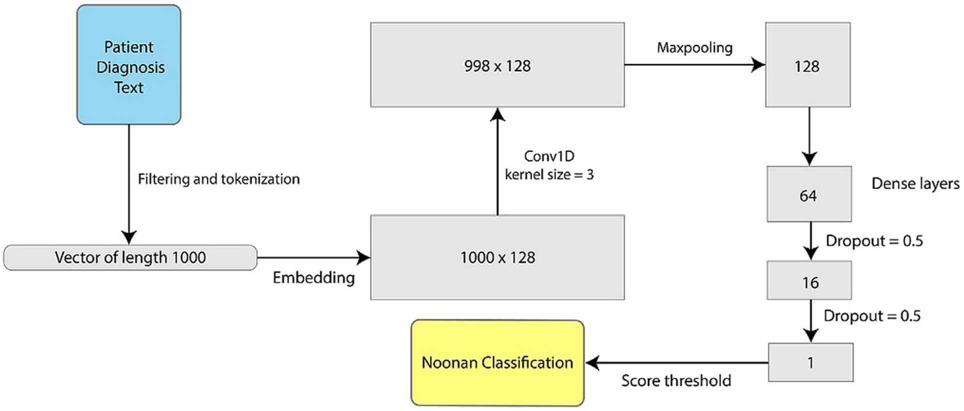

Figure 1 **Overview of the CNN model architecture.** After extracting the diagnosis descriptions for each patient, they were processed and tokenized as described in the method section. These padded vectors served as the inputs to the network, which began with a 128D embedding layer. The embedded words were passed to a 1-dimensional convolution layer with a window size of 3 words and a stride of 1. The output was max-pooled across the entire time dimension and then entered several fully connected layers. Dropout layers were used between the Dense layers for regularization. Conv 1D, 1 dimensional convolution.

text with the highest prediction score drop was ranked as the most informative feature for the patient's NS diagnosis.

Results

Cross-validation and training result

The PRAUC values in cross-validation over 20 epochs were plotted for all models (Supplemental Figure 1). The models began overfitting past a certain epoch, which was reflected by the decreasing validation PRAUC. The epoch with the highest average PRAUC during cross-validation was used as the number of epochs to train the final model. For instance, the CNN model achieved the highest average PRAUC at epoch 8. The final CNN model, therefore, used 8 epochs for training. The model using a CNN layer consistently achieved the best performance in the cross-validation. The architecture of the model is shown in Figure 1.

The fact that the CNN model performed better than MLP, which was anticipated, suggests that local patterns in diagnosis description text contained more information than individual terms. The CNN model also achieved higher PRAUC than the recurrent networks: LSTM, GRU, and their bidirectional variant models. This may indicate that most of the information in the diagnosis description texts used as features may be contained in local phrases rather than complex semantic relationships or time-sequential data, factors which normally would favor recurrent networks.

After training the final CNN model, we applied it to all 3719 unique diagnosis texts in the 189 NS cases. We found that the prediction scores of the diagnosis texts were highly negatively correlated with their P values from the word permutation test (Spearman's correlation P = 2.6e-267), suggesting that the higher scored diagnosis texts were more sensitive to changes in word order. The top 20 description texts with the highest prediction scores together with their frequencies in NS cases and all patients are listed in Table 1. As shown in the table, all these description texts had significant permutation P values. These terms included "short stature," "pulmonary valve stenosis," "ptosis," and "Cafe-au-lait spots," all of which are known common phenotypes associated with NS.[3,26] There are also terms related to abnormal partial thromboplastin time, gastrostomy, and abnormal electrocardiogram, which are not documented in reference genetic disease databases such as OMIM and HPO but have been observed in previous studies.[26-29] As shown in Table 1, all top 20 texts were more frequent in NS cases than the background population, suggesting their relevance to NS diagnosis.

Interestingly, the text with the highest score was "other specified congenital anomalies." Clinicians may use this code when they encounter a patient with suspected congenital anomalies at CCHMC. Although this does not map to any specific phenotype, it does appear to be highly enriched in patients with NS diagnosis. As expected, the terms "male" and "female" alone yielded very low prediction scores, suggesting gender was not informative for NS detection.

Another interesting observation was that there were many similar texts in the top 20 results. For example, although "partial thromboplastin time increased" was not very frequent among patients with NS, it is similar to "abnormal partial thromboplastin time" and "prolonged prothrombin time and partial thromboplastin time," and received a high prediction score. This probably reflected the nature of the CNN-based model, which considered the local patterns of words during learning and prediction.

Performance of NS detection in test set 1

The precision-recall curves for the models on test set 1 are shown in Figure 2. The CNN model again achieved the best performance in this test. The PRAUC of this model was

Table 1 Top scored diagnosis description by CNN model in NS classification

Rank	Text	Frequency in NS Patients, %	Frequency in All Patients, %	Odds Ratio	Permutation *P* value
1	Other specified congenital anomalies	54.497	0.601	206.6	.002
2	Abnormal partial thromboplastin time (PTT)	3.175	0.042	81.8	.001
3	Short stature associated with genetic disorder	10.582	0.063	204.6	.001
4	Prolonged prothrombin time (PT) and partial thromboplastin time (PTT)	1.058	0.004	338.7	.001
5	Partial thromboplastin time increased	0.529	0.001	1094.9	.003
6	Other congenital anomalies of pulmonary valve	8.466	0.102	94.1	.001
7	Short stature (child)	14.815	0.645	27.0	.006
8	Attention to gastrostomy	3.704	0.605	6.3	.001
9	Occlusion and stenosis of right carotid artery	0.529	0.006	91.2	.001
10	Abnormal electrocardiogram	1.058	0.136	7.9	.001
11	Nonrheumatic pulmonary valve stenosis with insufficiency	0.529	0.006	95.2	.001
12	Spitting up infant	0.529	0.167	3.2	.006
13	Observation for other specified suspected conditions	2.116	0.678	3.2	.001
14	Family history of congenital anomalies	0.529	0.074	7.2	.001
15	Ptosis, congenital, bilateral	0.529	0.002	273.8	.004
16	Congenital stenosis of pulmonary valve	30.159	0.370	120.8	.002
17	Advanced care planning/counseling discussion	1.058	0.022	50.0	.010
18	Pulmonary valve stenosis, unspecified etiology	8.995	0.046	237.9	.003
19	Cafe-au lait spots	1.058	0.064	16.7	.001
20	Vasovagal syncope	0.529	0.335	1.6	.006

Each diagnosis description from the data set of patients with NS was evaluated by the final CNN model. This table shows 20 of the diagnosis description terms with the highest prediction score. The frequencies of these terms in NS cases and all patients and their odds ratios are included. The permutation *P* value was derived by performing one-sided one-sample Wilcox rank-sum test of the prediction scores of the permuted diagnosis texts against the score of the original diagnosis text. Several description terms that are different spellings of the same diagnosis were omitted to reduce redundancy. This table helps show the clinical factors used by the model to make a classification.

CNN, convolution neural net; *NS*, Noonan syndrome.

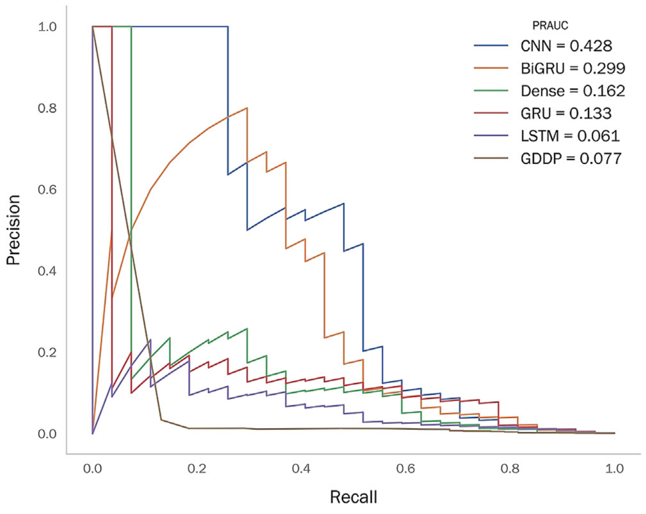

Figure 2 **Precision-recall curves of test set 1 across different models.** The precision, recall, and PRAUC for each model using test set 1. The tested deep learning architectures included CNN, LSTM, GRU, BiGRU, and Dense (MLP) networks. The performance of the previously developed method GDDP is also displayed. The CNN model had the highest PRAUC of 0.428, with 0.47 precision at 0.52 recall. BiGRU, bidirectional gated-recurrent unit; CNN, convolution neural net; GDDP, Genetic Disease Diagnosis based on Phenotypes; GRU, gated-recurrent unit; LSTM, long short-term memory; MLP, multilayer perceptron; PRAUC, area under the precision-recall curve.

0.43. With a score cutoff of 0.84, 30 positive patients were discovered, 14 of which had existing NS diagnosis, corresponding to a precision 0.47 with a recall of 0.52, or a sensitivity 0.52 with a specificity of 1.00. For reference, a random model had 0.001 precision at 0.5 recall.

We applied GDDP, our previously developed computational method, to test set 1. Because GDDP prioritizes genetic diseases based on HPO terms, the diagnosis description texts were first converted to HPO terms as described before.[8] There were multiple subtypes of NS in OMIM. For each test patient, the minimum rank of all of these NS subtypes was used as the rank of NS by GDDP. The precision and recall were calculated based on the rank of NS for that patient.[8] As shown in Figure 1, the PRAUC of GDDP was 0.077. All deep learning models, except LSTM, performed significantly better than GDDP, and the CNN model was 35 times more precise than GDDP.

Pulmonary stenosis and short stature are the 2 most representative phenotypes associated with NS. For comparison purposes, we also tried patient screening based on these 2 phenotypes. Using pulmonary stenosis alone (patients whose diagnosis text containing “pulmonary valve stenosis,” “pulmonary stenosis,” or “stenosis of pulmonary valve”), the precision and recall were 0.068 and 0.37, respectively. Using short stature alone, the precision and recall were 0.056 and 0.74, respectively. When identifying patients with both pulmonary stenosis and short stature, the

Table 2 Summary of test results

Test Set 1								
Method	Total	NS	Positive	TP	Precision	Recall	PRAUC	F1
PS	27027	27	148	10	0.068	0.37	–	0.115
SS	27027	27	355	20	0.056	0.74	–	0.104
PS+SS	27027	27	13	9	0.69	0.33	–	0.446
CNN	27027	27	39	14	0.47	0.52	**0.428**	**0.494**
GDDP	27027	27	451	6	0.013	0.22	0.077	0.025
Test Set 2								
Method	Total	NS	Positive	TP	Precision	Recall	PRAUC	F1
PS	10010	10	69	6	0.087	0.60	–	0.152
SS	10010	10	252	3	0.012	0.30	–	0.023
PS+SS	10010	10	1	1	1	0.10	–	0.182
CNN	10010	10	12	4	0.33	0.40	**0.160**	**0.362**
GDDP	10010	10	289	3	0.01	0.30	0.009	0.019

Results of NS detection using different methods in test sets 1 and 2. The first 4 columns display the count for total patients in the test set (total), NS cases in the test set (NS), all patients identified by the method (positive), and true NS cases in the identified patients (TP). The last 4 columns are the precision, recall, PRAUC (if available), and F1 score of the method. The phenotype-based method for identifying patients with NS searches diagnosis description texts for the most common terms associated with NS; these terms included pulmonary stenosis (PS), short stature (SS), and both pulmonary stenosis and short stature (PS+SS). The CNN-based model achieved the highest F1 scores in both tests. F1 score was defined as the harmonic mean of the precision and recall. Boldfaced values indicate the test results of the CNN model.

CNN, convolution neural net; *GDDP*, Genetic Disease Diagnosis based on Phenotypes; *NS*, Noonan syndrome; *PRAUC*, area under the precision-recall curve.

precision and recall were 0.69 and 0.33, respectively. The test results are listed in Table 2. As can be seen in the table, in terms of F1 score, the CNN model performed better than the phenotype-based approaches, which were better than GDDP.

Performance of NS detection in test set 2

To further evaluate the CNN model's performance, we applied it to test set 2. This test set contained 10 undiagnosed NS cases (when the EHR data were extracted) and 10,000 random controls. The precision-recall curve of the CNN model is shown in Figure 3. The PRAUC was 0.160. With a score cutoff of 0.84, 12 positive patients were discovered, 4 of which were the undiagnosed patients with NS, corresponding to a precision 0.33 with a recall of 0.40, or a sensitivity 0.40 with a specificity of 1.00. In contrast, the PRAUC of GDDP on the same test set was 0.009. With rank 10 as cutoff, GDDP discovered 289 positive patients, 3 of which were the undiagnosed patients with NS, corresponding to a precision 0.01 with a recall of 0.30.

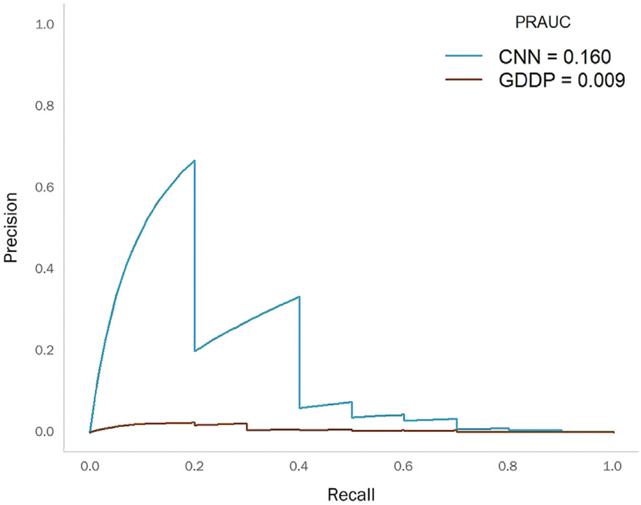

Figure 3 Precision-recall curves of CNN model and GDDP on test set 2. The precision, recall, and PRAUC for CNN model and GDDP using test set 2. The PRAUC for the CNN model and GDDP were 0.16 and 0.009, respectively. CNN, convolution neural net; GDDP, Genetic Disease Diagnosis based on Phenotypes; PRAUC, area under the precision-recall curve.

We repeated the phenotype-based patient screening on test set 2. Using pulmonary stenosis alone, the precision and recall were 0.087 and 0.60, respectively. Using short stature alone, the precision and recall were 0.012 and 0.30, respectively. When identifying patients with both pulmonary stenosis and short stature, the precision and recall were 1 and 0.10, respectively. The test results are listed in Table 2. The performance of the CNN model was again better than the simple phenotype-based approaches.

Web application

A Shiny app (http://eNS.research.cchmc.org) was developed to show how the pretrained deep learning model can be used to score a patient for NS based on the diagnosis description texts, and how an explanation can be derived for the patient. As shown in the screenshot (Supplemental Figure 2), the diagnosis texts of a patient were entered in the text area in the left panel and tokenized and transformed as input to the pretrained CNN model. The NS score of the patient generated by the model would then be displayed in the main panel. The sensitivity analysis result is shown in the table below, ordered by the prediction score drop when the text is removed from the patient. As displayed, "congenital pulmonary valve stenosis" and "short stature (child)" were the most informative texts for this patient's NS classification according to the CNN model.

Discussion

Clinician recognition of characteristic symptoms and subsequent referrals for testing is the most common path to diagnosis for patients with genetic diseases. However, certain conditions, such as NS, are often difficult for pediatricians and primary care physicians to recognize, thus creating a need for improvements to the diagnostic process. In this study, we developed a computational method to detect NS in a pediatric population based on EHR diagnosis texts and deep learning. We tested our approach using 2 independent test sets containing diagnosed and undiagnosed NS cases, respectively, and our CNN-based model showed improved performance over the simple phenotype-based approaches and our previous method. We argue that these prediction scores have the potential to improve diagnosis and treatment by helping clinicians prioritize referrals for patients with a high risk of genetic disorders.

Our method offers several features that help resolve key issues in a population-scale computational screening of rare diseases: (1) the approach used only the diagnosis description texts, data which are readily available for a large pediatric population, (2) the deep learning-based models do not require explicit natural language processing and manual feature selection, which reduced the complexity of data processing and model development, (3) with diagnosis description texts directly as input, our method could efficiently perform large-scale systematic patient screening based on EHR data, (4) our method could provide an interpretation to highlight the texts that are most informative for the decision making, providing clinical insights that could reduce validation times, and (5) even with a relatively small set of NS cases for training, our EHR-based method showed clinically actionable performance, making it applicable for both NS and other rare genetic disorders.

This last point is especially important and valuable. There are more than 7000 rare genetic diseases, collectively affecting 8% of people in the United States.[30] Many patients have to experience years long diagnostic odyssey to confirm the diagnosis owing to lack of recognition and referral.[31,32] As suggested by recent literature[7] and this study, there could be a substantial proportion of people with unrecognized NS in the general population. With a method for systematic screening, some of these patients could be recognized and referred earlier to clinical geneticists, thus shortening time to diagnosis for those individuals.

The potential for this was shown by the 2 independent tests in the study, particularly test set 2, which contained cases that were not diagnosed by the time the EHR data were exported. Although we could identify some NS cases based on the presence of either pulmonary stenosis or short stature in their EHR, the precision was very low. Selecting patients with both phenotypes resulted in higher precision but much lower recall. We also previously developed GDDP to guide the identification of pathogenic variants for individual patients. GDDP had low precision and recall in our tests, suggesting that it is not ideal to use GDDP to screen patients with NS. Our CNN-based deep learning model struck a strong balance between precision and recall and had the highest F1 scores among all methods in both tests. After taking a closer look at the 6 patients in test set 2 who were not identified by the CNN model (summarized in Supplemental Table 1), we found that 2 of the patients did not have any typical NS phenotypes recorded in their diagnosis texts. The third patient had short stature and developmental delay but no cardiac abnormalities. Because these patients did not contain informative phenotypes in their diagnosis texts, the chance of recognizing them using diagnosis texts was low. The 4 patients who were identified by the CNN model all had pulmonary stenosis and other phenotypes related to NS, such as short stature or atrial septal defect. Although we do not know why they were not diagnosed by 2019, they may have been identified earlier if our approach had been clinically used in 2019.

There have been several previous studies that used facial images to detect NS. Tekendo-Ngongang and Kruszka[12] applied the DeepGestalt model[13] to differentiate patients with NS from unaffected controls on the basis of facial images in a cohort of African individuals. More recently Yang et al[11] developed facial recognition models to separate patients with NS from healthy children and children with other dysmorphic syndromes. Both studies reported high accuracies with the area under the receiver operating characteristic curve higher than 0.9 in test sets with relatively high case to control ratios (close to 1:1). However, given that these facial image-based methods were designed and tested for small-scale scenarios in which the patient has already been suspected of a genetic disorder and our EHR texts-based method was designed for screening on a much larger scale, we anticipate that the 2 approaches will work complementarily.

There are several important limitations to this study that should be noted and explored further. First is the biases inherent to learning from EHR data. Some diagnosis terms in the records of patient with NS may be only present after genetic confirmation or physician suspicion of NS, leading to uncertainty regarding the ability of such models to detect undiagnosed and unsuspected patients. To mitigate these factors, we identified and removed the obvious confounding terms before training and testing. However, there may still be patterns such as comorbidities in the data unique to a diagnosis of NS rather than the physiological signs of NS. Second, models trained using data from a single institution may not generalize well to hospitals with clinicians who use a different set of EHR diagnosis descriptions. Third, with using diagnosed patients with NS as positive cases in the training set, the deep learning models are not optimized for recognizing patients without typical phenotypes. One recent study suggested a similar finding as observed in our test set 2, the undiagnosed patients with NS might not manifest typical phenotypes, such as short stature and heart anomalies.[7] This limitation is intrinsic to the supervised learning strategy, which can be alleviated by including more diversified cases in the training set. Finally, there is still a potential to improve the model performance. The number of NS cases for training can be increased by pooling more cases from multiple pediatric hospitals. At the same time, additional machine-learning strategies such as multitask learning[33,34] may be able to boost single task performance of the models by learning common underlying principles from diseases with features similar to NS, such as Alagille syndrome, Williams syndrome, and CHARGE syndrome.

Conclusion

In conclusion, we showed that deep learning models based on diagnosis description texts from EHR are viable for detecting children with Noonan syndrome. The CNN-based diagnostic model outperformed the method based on searching of key phenotypes and our previous statistical method based on comparison of HPO terms. In addition, our method provides a clinically relevant evaluation of the diagnostic values of the diagnosis texts of a patient. This study proposes an efficient way to screen for undiagnosed cases of NS and other diseases from a children's hospital's EHR database.

Data Availability

The raw data of this study are available on request to the corresponding author. The data release outside Cincinnati Children's Hospital Medical Center will require approval from Cincinnati Children's Hospital Medical Center legal department. The data are not publicly available owing to privacy or ethical restrictions. All deep learning models were implemented using TensorFlow v.2.5.0 and Keras v2.5.0-tf. All source code is available on GitHub (https://github.com/xiaojoey/eNS).

Acknowledgments

A portion of the background for this research was obtained through PEDSnet, a Pediatric Clinical Research Network. PEDSnet's participation in The National Patient-Centered Clinical Research Network (PCORnet) is funded through The Patient-Centered Outcomes Research Institute (PCORI) Award RI-CHOP-01-PS1. The authors would like to acknowledge Diana Epperson and Janet Zahner for helping to query and organize data from the Cincinnati Children's Hospital Medical Center Informatics for Integrating Biology and the Bedside database. The authors would also like to acknowledge Nathan Pajor for helping with PEDSnet database queries. This study was partially funded by the Center for Pediatric Genomics, Cincinnati Children's Hospital Medical Center.

Author Information

Conceptualization: J.C., K.N.W.; Data Curation: Z.Y., J.C.; Formal Analysis: Z.Y., J.C.; Investigation: Z.Y., J.C.; Methodology: Z.Y., J.C., Y.N.; Software: Z.Y., J.C.; Writing-original draft: Z.Y., J.C.; Writing-review and editing: Z.Y., A.S., Y.N., G.Z., K.N.W., J.C.

Ethics Declaration

This study protocol was reviewed and approved by the Cincinnati Children's Hospital Institutional Review Board (protocol number 2020-0685).

Conflict of Interest

The authors declare no conflict of interest.

Additional Information

The online version of this article (https://doi.org/10.1016/j.gim.2022.08.002) contains supplementary material, which is available to authorized users.

References

1. Prendiville TW, Gauvreau K, Tworog-Dube E, et al. Cardiovascular disease in Noonan syndrome. *Arch Dis Child*. 2014;99(7):629-634. http://doi.org/10.1136/archdischild-2013-305047
2. Romano AA, Allanson JE, Dahlgren J, et al. Noonan syndrome: clinical features, diagnosis, and management guidelines. *Pediatrics*. 2010;126(4):746-759. http://doi.org/10.1542/peds.2009-3207
3. Roberts AE, Allanson JE, Tartaglia M, Gelb BD. Noonan syndrome. *Lancet*. 2013;381(9863):333-342. http://doi.org/10.1016/S0140-6736(12)61023-X
4. Chen PC, Yin J, Yu HW, et al. Next-generation sequencing identifies rare variants associated with Noonan syndrome. *Proc Natl Acad Sci U S A*. 2014;111(31):11473-11478. http://doi.org/10.1073/pnas.1324128111
5. Tafazoli A, Eshraghi P, Koleti ZK, Abbaszadegan M. Noonan syndrome – a new survey. *Arch Med Sci*. 2017;13(1):215-222. http://doi.org/10.5114/aoms.2017.64720
6. Forrest CB, Margolis PA, Bailey LC, et al. PEDSnet: a national pediatric learning health system. *J Am Med Inform Assoc*. 2014;21(4):602-606. http://doi.org/10.1136/amiajnl-2014-002743
7. Wenger BM, Patel N, Lui M, et al. A genotype-first approach to exploring Mendelian cardiovascular traits with clear external

manifestations. *Genet Med*. 2021;23(1):94-102. http://doi.org/10.1038/s41436-020-00973-2

8. Chen J, Xu H, Jegga A, Zhang K, White PS, Zhang G. Novel phenotype-disease matching tool for rare genetic diseases. *Genet Med*. 2019;21(2):339-346. http://doi.org/10.1038/s41436-018-0050-4
9. Köhler S, Gargano M, Matentzoglu N, et al. The Human Phenotype Ontology in 2021. *Nucleic Acids Res*. 2021;49(D1):D1207-D1217. http://doi.org/10.1093/nar/gkaa1043
10. Hamosh A, Scott AF, Amberger JS, et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*. 2005;33(Database issue):D514-D517. http://doi.org/10.1093/nar/gki033
11. Yang H, Hu XR, Sun L, et al. Automated facial recognition for Noonan syndrome using novel deep convolutional neural network with additive angular margin loss. *Front Genet*. 2021;12:669841. http://doi.org/10.3389/fgene.2021.669841
12. Tekendo-Ngongang C, Kruszka P. Noonan syndrome on the African Continent. *Birth Defects Res*. 2020;112(10):718-724. http://doi.org/10.1002/bdr2.1675
13. Gurovich Y, Hanani Y, Bar O, et al. Identifying facial phenotypes of genetic disorders using deep learning. *Nat Med*. 2019;25(1):60-64. http://doi.org/10.1038/s41591-018-0279-0
14. World Health Organization. *International statistical classification of diseases and related health problems*. World Health Organization; 2016. 10th revision, Fifth edition. Accessed August 29, 2022. https://apps.who.int/iris/handle/10665/246208
15. Burrows EK, Razzaghi H, Utidjian L, Bailey LC. Standardizing clinical diagnoses: evaluating alternate terminology selection. *AMIA Jt Summits Transl Sci Proc*. 2020;2020:71-79.
16. Anderson K, Cnota J, James J, et al. Prevalence of Noonan spectrum disorders in a pediatric population with valvar pulmonary stenosis. *Congenit Heart Dis*. 2019;14(2):264-273. http://doi.org/10.1111/chd.12721
17. 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(3):e0118432. http://doi.org/10.1371/journal.pone.0118432
18. Rumelhart DE, Hinton GE, Williams RJ. *Learning Internal Representations by Error Propagation*. California Univ San Diego La Jolla Inst for Cognitive Science; 1985.
19. LeCun Y, Bengio Y, GJn Hinton. Deep learning. *Nature*. 2015;521(7553):436-444. http://doi.org/10.1038/nature14539
20. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9(8):1735-1780. http://doi.org/10.1162/neco.1997.9.8.1735
21. Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics; 2014:1724-1734.
22. Ciregan D, Meier U, Schmidhuber J. Multi-column deep neural networks for image classification. In: 2012 *IEEE conference on computer vision and pattern recognition*. IEEE; 2012:3642-3649.
23. Ching T, Himmelstein DS, Beaulieu-Jones BK, et al. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface*. 2018;15(141), 20170387. http://doi.org/10.1098/rsif.2017.0387
24. Pennington J, Socher R, Manning C. Glove: global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics; 2014:1532-1543.
25. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery; 2016:1135-1144.
26. Tartaglia M, Gelb BD, Zenker M. Noonan syndrome and clinically related disorders. *Best Pract Res Clin Endocrinol Metab*. 2011;25(1):161-179. http://doi.org/10.1016/j.beem.2010.09.002
27. Artoni A, Selicorni A, Passamonti SM, et al. Hemostatic abnormalities in Noonan syndrome. *Pediatrics*. 2014;133(5):e1299-e1304. http://doi.org/10.1542/peds.2013-3251
28. Nugent DJ, Romano AA, Sabharwal S, Cooper DL. Evaluation of bleeding disorders in patients with Noonan syndrome: a systematic review. *J Blood Med*. 2018;9:185-192. http://doi.org/10.2147/JBM.S164474
29. Croonen EA, van der Burgt I, Kapusta L, Draaisma JM. Electrocardiography in Noonan syndrome PTPN11 gene mutation—phenotype characterization. *Am J Med Genet A*. 2008;146A(3):350-353. http://doi.org/10.1002/ajmg.a.32140
30. Wakap SN, Lambert DM, Olry A, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur J Hum Genet*. 2020;28(2):165-173. http://doi.org/10.1038/s41431-019-0508-0
31. Gainotti S, Mascalzoni D, Bros-Facer V, et al. Meeting patients' right to the correct diagnosis: ongoing international initiatives on undiagnosed rare diseases and ethical and social issues. *Int J Environ Res Public Health*. 2018;15(10):2072. http://doi.org/10.3390/ijerph15102072
32. Wise AL, Manolio TA, Mensah GA, et al. Genomic medicine for undiagnosed diseases. *Lancet*. 2019;394(10197):533-540. http://doi.org/10.1016/S0140-6736(19)31274-7
33. Zhang Z, Luo P, Loy CC, Tang X. Facial landmark detection by deep multi-task learning. In: *European conference on computer vision*. Springer; 2014.
34. Lu Y, Kumar A, Zhai S, Cheng Y, Javidi T, Feris R. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In: *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2017:1131-1140.