nature medicine

LETTERS

https://doi.org/10.1038/s41591-018-0335-9

Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence

Huiying Liang[1,8], Brian Y. Tsui[2,8], Hao Ni[3,8], Carolina C. S. Valentim[4,8], Sally L. Baxter[2,8], Guangjian Liu[1,8], Wenjia Cai[2], Daniel S. Kermany[1,2], Xin Sun[1], Jiancong Chen[2], Liya He[1], Jie Zhu[1], Pin Tian[2], Hua Shao[2], Lianghong Zheng[5,6], Rui Hou[5,6], Sierra Hewett[1,2], Gen Li[1,2], Ping Liang[3], Xuan Zang[3], Zhiqi Zhang[3], Liyan Pan[1], Huimin Cai[5,6], Rujuan Ling[1], Shuhua Li[1], Yongwang Cui[1], Shusheng Tang[1], Hong Ye[1], Xiaoyan Huang[1], Waner He[1], Wenqing Liang[1], Qing Zhang[1], Jianmin Jiang[1], Wei Yu[1], Jianqun Gao[1], Wanxing Ou[1], Yingmin Deng[1], Qiaozhen Hou[1], Bei Wang[1], Cuichan Yao[1], Yan Liang[1], Shu Zhang[1], Yaou Duan[2], Runze Zhang[2], Sarah Gibson[2], Charlotte L. Zhang[2], Oulan Li[2], Edward D. Zhang[2], Gabriel Karin[2], Nathan Nguyen[2], Xiaokang Wu[1,2], Cindy Wen[2], Jie Xu[2], Wenqin Xu[2], Bochu Wang[2], Winston Wang[2], Jing Li[1,2], Bianca Pizzato[2], Caroline Bao[2], Daoman Xiang[1], Wanting He[1,2], Suiqin He[2], Yugui Zhou[1,2], Weldon Haw[2,7], Michael Goldbaum[2], Adriana Tremoulet[2], Chun-Nan Hsu[2], Hannah Carter[2], Long Zhu[3], Kang Zhang[1,2,7*] and Huimin Xia[1*]

Artificial intelligence (AI)-based methods have emerged as powerful tools to transform medical care. Although machine learning classifiers (MLCs) have already demonstrated strong performance in image-based diagnoses, analysis of diverse and massive electronic health record (EHR) data remains challenging. Here, we show that MLCs can query EHRs in a manner similar to the hypothetico-deductive reasoning used by physicians and unearth associations that previous statistical methods have not found. Our model applies an automated natural language processing system using deep learning techniques to extract clinically relevant information from EHRs. In total, 101.6 million data points from 1,362,559 pediatric patient visits presenting to a major referral center were analyzed to train and validate the framework. Our model demonstrates high diagnostic accuracy across multiple organ systems and is comparable to experienced pediatricians in diagnosing common childhood diseases. Our study provides a proof of concept for implementing an AI-based system as a means to aid physicians in tackling large amounts of data, augmenting diagnostic evaluations, and to provide clinical decision support in cases of diagnostic uncertainty or complexity. Although this impact may be most evident in areas where healthcare providers are in relative shortage, the benefits of such an AI system are likely to be universal.

Medical information has become increasingly complex over time. The range of disease entities, diagnostic testing and biomarkers, and treatment modalities has increased exponentially in recent years. Subsequently, clinical decision-making has also become more complex and demands the synthesis of decisions from assessment of large volumes of data representing clinical information. In the current digital age, the electronic health record (EHR) represents a massive repository of electronic data points representing a diverse array of clinical information[1–3]. Artificial intelligence (AI) methods have emerged as potentially powerful tools to mine EHR data to aid in disease diagnosis and management, mimicking and perhaps even augmenting the clinical decision-making of human physicians[1].

To formulate a diagnosis for any given patient, physicians frequently use hypotheticodeductive reasoning. Starting with the chief complaint, the physician then asks appropriately targeted questions relating to that complaint. From this initial small feature set, the physician forms a differential diagnosis and decides what features (historical questions, physical exam findings, laboratory testing, and/or imaging studies) to obtain next in order to rule in or rule out the diagnoses in the differential diagnosis set. The most useful features are identified, such that when the probability of one of the diagnoses reaches a predetermined level of acceptability, the process is stopped, and the diagnosis is accepted. It may be possible to achieve an acceptable level of certainty of the diagnosis with only a few features without having to process the entire feature set. Therefore, the physician can be considered a classifier of sorts.

In this study, we designed an AI-based system using machine learning to extract clinically relevant features from EHR notes to mimic the clinical reasoning of human physicians. In medicine, machine learning methods have already demonstrated strong performance in image-based diagnoses, notably in radiology[2], dermatology[4], and ophthalmology[5–8], but analysis of EHR data presents a number of difficult challenges. These challenges include the vast quantity of data, high dimensionality, data sparsity, and deviations

[1]Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, China. [2]Institute for Genomic Medicine, Institute of Engineering in Medicine, and Shiley Eye Institute, University of California, San Diego, La Jolla, CA, USA. [3]Hangzhou YITU Healthcare Technology Co. Ltd, Hangzhou, China. [4]Department of Thoracic Surgery/Oncology, First Affiliated Hospital of Guangzhou Medical University, China State Key Laboratory and National Clinical Research Center for Respiratory Disease, Guangzhou, China. [5]Guangzhou Kangrui Co. Ltd, Guangzhou, China. [6]Guangzhou Regenerative Medicine and Health Guangdong Laboratory, Guangzhou, China. [7]Veterans Administration Healthcare System, San Diego, CA, USA. [8]These authors contributed equally: Huiying Liang, Brian Tsui, Hao Ni, Carolina C. S. Valentim, Sally L. Baxter, Guangjian Liu. *e-mail: kang.zhang@gmail.com; xiahumin@hotmail.com

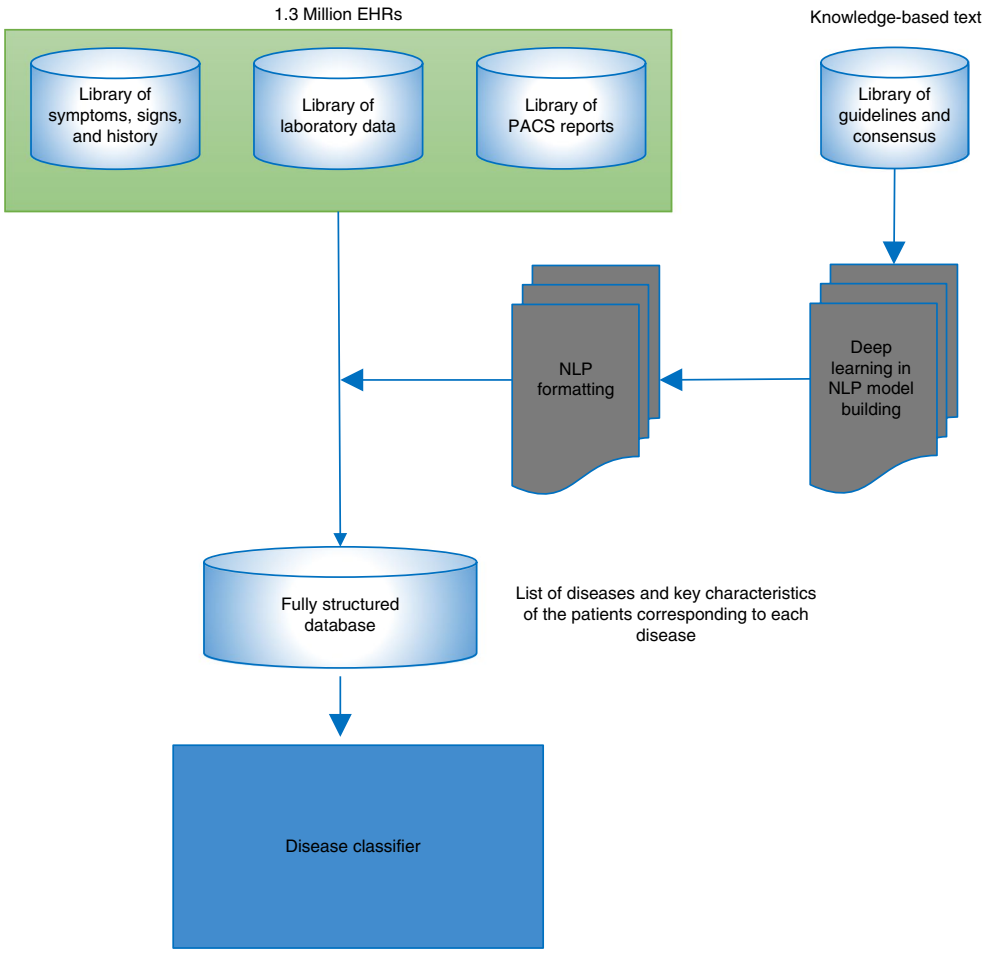


Fig. 1 | Workflow diagram of our AI pediatric diagnosis framework. This diagram depicts the process of data extraction from electronic medical records, followed by deep learning-based NLP analysis of these encounters, which were then processed with a disease classifier to predict a clinical diagnosis for each encounter.

or systematic errors in medical data[9]. These challenges make it difficult to use machine learning methods to perform accurate pattern recognition and generate predictive clinical models.

In this paper, we propose a data mining framework for EHR data that integrates prior medical knowledge and data-driven modeling. We develop a deep learning-based natural language processing (NLP) system to extract clinically relevant information and subsequently establish a diagnostic system based on extracted clinical features. Finally, this framework is applied in a large pediatric population to demonstrate the diagnostic ability of an AI-based method.

We conducted a retrospective study and obtained EHRs from 1,362,559 outpatient visits from 567,498 patients of the Guangzhou Women and Children's Medical Center, a major academic medical referral center. These records encompassed physician–patient encounters presenting from January 2016 to July 2017. The median age was 2.35 years (range 0 to 18 years; 95% confidence interval 0.2 to 9.7 years), and 40.11% were female (Supplementary Table 1).

The primary diagnoses considered 55 diagnosis codes in total, encompassing common pediatric diseases and representing a wide range of pathologies. Some of the most frequently encountered diagnoses included acute upper respiratory infection, bronchitis, diarrhea, bronchopneumonia, acute tonsillitis, stomatitis, and acute sinusitis (Supplementary Table 1). The records originated from a wide range of specialties, with the top three most represented departments being general pediatrics, the Special Clinic for Children, and pediatric pulmonology (Supplementary Table 1). The Special Clinic for Children is for private patients at this institution and encompassed care for a range of conditions.

First, the diagnostic system analyzed the EHR in the absence of a defined classification system with human input. In the absence of pre-defined labeling as input, the unsupervised clustering was still able to detect trends in clinical features to generate a relatively sensible grouping structure (Extended Data 1). In many instances, it successfully established broad grouping of related diagnoses even without any directed labeling or classification system in place, suggesting that the clinical features that we developed capture the key similarities and differences between the conditions that we intend to model and diagnose.

A total of 6,183 charts were manually annotated using the schema described in the Methods section by senior attending physicians with more than 25 years' clinical practice experience. Then 3,564 manually annotated charts were used to train the NLP information extraction model, and the remaining 2,619 were used to validate the model. The information extraction model summarized the key conceptual categories representing clinical data (Fig. 1). This NLP model utilized deep learning techniques (see Methods) to automate the annotation of the free text EHR notes into the standardized lexicon and clinical features, allowing the further processing of clinical information for diagnostic classification.

The NLP model achieved excellent results in the annotation of EHR physician notes (Supplementary Table 2). Across all categories of clinical data (chief complaint, history of present illness, physical

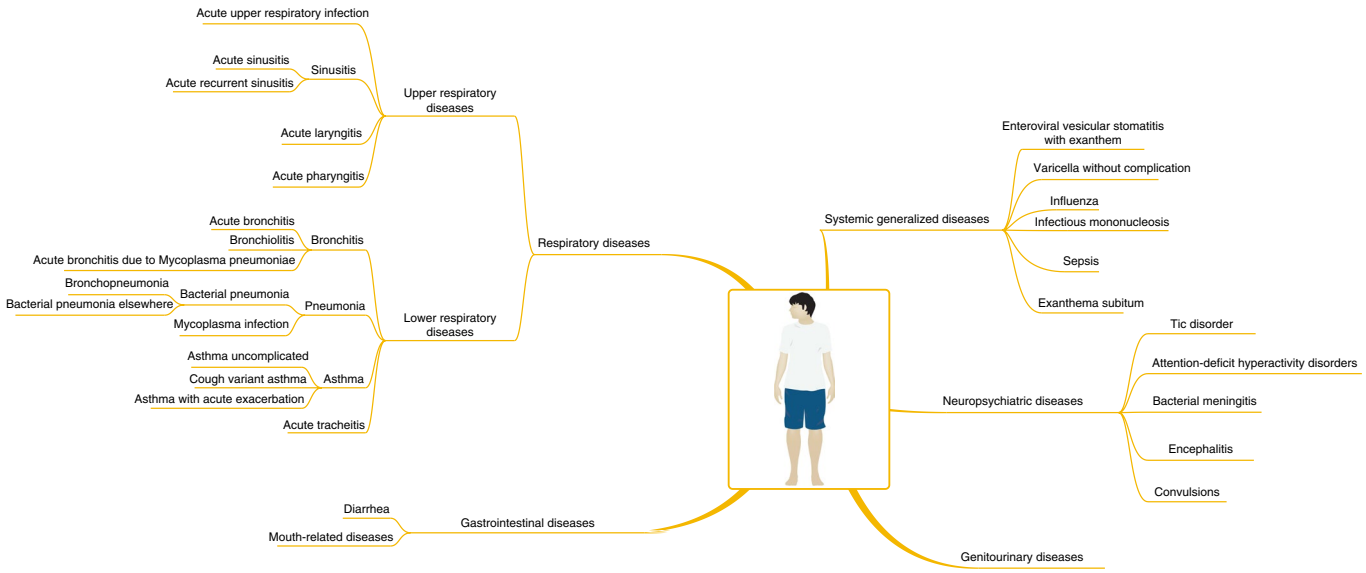


Fig. 2 | Hierarchy of the diagnostic framework in a large pediatric cohort. A hierarchical logistic regression classifier was used to establish a diagnostic system based on anatomic divisions. An organ-based approach was used, wherein diagnoses were first separated into broad organ systems, then subsequently divided into organ subsystems and/or into more specific diagnosis groups.

examination, laboratory testing, and PACS (picture archiving and communication systems) reports), the F1 scores exceeded 90% except in one instance, which was for categorical variables detected in the laboratory testing. The highest recall of the NLP model was achieved for physical examination (95.62% for categorical variables, 99.08% for free text), and the lowest for laboratory testing (72.26% for categorical variables, 88.26% for free text). The highest precision of the NLP model was for chief complaint (97.66% for categorical variables, 98.71% for free text), and the lowest for laboratory testing (93.78% for categorical variables, and 96.67% for free text). In general, the precision (or positive predictive value) of the NLP labeling was slightly greater than the recall (the sensitivity), but the system demonstrated overall strong performance across all domains (Supplementary Table 2).

After the EHR notes were annotated using the deep NLP information extraction model, logistic regression classifiers were used to establish a hierarchical diagnostic system. The diagnostic system was primarily based on anatomic divisions (for example, organ systems). This was meant to mimic traditional frameworks used in physician reasoning, in which an organ-based approach can be employed for the formulation of a differential diagnosis. Logistic regression classifiers were used to allow straightforward identification of relevant clinical features and for ease of establishing interpretability for the diagnostic classification.

The first level of the diagnostic system categorized the EHR notes into broad organ systems: respiratory, gastrointestinal, neuropsychiatric, genitourinary, and systemic or generalized conditions (Fig. 2). This was the first level of separation in the diagnostic hierarchy. Then, within each organ system, further sub-classifications and hierarchical layers were made, where applicable. The most number of diagnoses in this cohort fell into the respiratory system, which was further divided into upper respiratory conditions and lower respiratory conditions. These were further separated into more specific anatomic divisions (for example, laryngitis, tracheitis, bronchitis, and pneumonia) (see Methods). The performance of the classifier was evaluated at each level of the diagnostic hierarchy. In short, the system was designed to evaluate the extracted features of each patient record and categorize the set of features into finer levels of diagnostic specificity along the levels of this decision tree, similar to how a human physician might evaluate a patient's features to achieve a diagnosis based on the same clinical data incorporated into the information model. Encounters labeled by physicians as having a primary diagnosis of 'fever' or 'cough' were eliminated, as these represented symptoms rather than specific disease entities.

Across all levels of the diagnostic hierarchy, our diagnostic system achieved a high level of accuracy between the predicted primary diagnoses based on the extracted clinical features by the NLP information model and the initial diagnoses designated by the examining physician (Table 1). For the first level, in whichthe diagnostic system classified the patient's diagnosis into a broad organ system, the median accuracy was 0.90, ranging from 0.85 for gastrointestinal diseases to 0.98 for neuropsychiatric disorders (full contingency table in Table 1a). Even at deeper levels of diagnostic specification, the system retained a strong level of performance. To illustrate, within the respiratory system, the next division in the diagnostic hierarchy was between upper respiratory and lower respiratory conditions. The system achieved an accuracy of 0.89 of upper respiratory conditions and 0.87 of lower respiratory conditions between predicted diagnoses and initial diagnoses (Table 1b). When dividing the upper respiratory subsystem into more specific categories, the median accuracy was 0.92 (range: 0.86 for acute laryngitis to 0.96 for sinusitis, Table 1c). Acute upper respiratory infection was the single most common diagnosis among the cohort, and our model was able to accurately predict the diagnosis in 95% of the encounters (Table 1c). Within the respiratory system, asthma was categorized separately as its own subcategory, and the accuracy ranged from 0.83 for cough variant asthma to 0.97 for unspecified asthma with acute exacerbation (Table 1d).

In addition to the strong performance in the respiratory system, the diagnostic model performed comparably in the other organ subsystems (see Supplementary Tables 3–6). Notably, the classifier achieved a very high level of accuracy in predicting diagnoses for the generalized systemic conditions (Supplementary Table 6), with an accuracy of 0.90 for infectious mononucleosis, 0.93 for roseola (sixth disease), 0.94 for influenza, 0.93 for varicella, and 0.97 for hand-foot-mouth disease. The diagnostic framework also achieved high accuracy for conditions with potential for high morbidity, such

Table 1 | Illustration of diagnostic performance of the logistic regression classifier at multiple levels of the diagnostic hierarchy

Physician-assigned diagnoses																	
Organ systems						**Respiratory system**			**Upper respiratory system**				**Asthma**				
	Resp. (*n* = 315,661)	Gast. (*n* = 41,098)	Sys. (*n* = 11,698)	Neuro (*n* = 8,410)	Geni. (*n* = 1,326)		U.Resp. (*n* = 156,176)	L.Resp. (*n* = 159,485)		AURI (*n* = 144,503)	Sin. (*n* = 8,828)	A. La. (*n* = 2845)		U.A. (*n* = 776)	CVA (*n* = 201)	AAE (*n* = 121)
Resp. (*n* = 295,403)	**0.920**	0.100	0.048	0.005	0.049	U. Resp. (*n* = 158,890)	**0.890**	0.110	AURI (*n* = 137,995)	**0.950**	0.033	0.110	U.A. (*n* = 740)	**0.910**	0.160	0.000
Gast. (*n* = 55,704)	0.063	**0.850**	0.066	0.005	0.044	L. Resp. (*n* = 156,771)	0.130	**0.870**	Sin. (*n* = 10,859)	0.016	**0.960**	0.028	CVA (*n* = 236)	0.085	**0.830**	0.033
Sys. (*n* = 14,267)	0.009	0.028	**0.870**	0.003	0.012				A. La. (*n* = 7,322)	0.033	0.010	**0.860**	AAE (*n* = 122)	0.004	0.010	**0.970**
Neuro. (*n* = 9,007)	0.002	0.003	0.003	**0.980**	0.005											
Geni. (*n* = 3,812)	0.006	0.014	0.008	0.004	**0.890**											

At the first level of the diagnostic hierarchy, the classifier accurately discerned broad anatomic classifications between organ systems in this large cohort of pediatric patients. For example, among 315,661 encounters with primary respiratory diagnoses as determined by human physicians, the deep learning-based model was able to correctly predict the diagnoses in 295,403 (92%) cases, resulting in an accuracy of 0.920 as shown in the table. Within the respiratory system, at the next level of the diagnostic hierarchy, the classifier could discern between upper respiratory conditions and lower respiratory conditions. Within the upper respiratory system, further distinctions could be made into acute upper respiratory infection, sinusitis, and laryngitis. Acute upper respiratory infection and sinusitis were among the most common conditions in the entire cohort, and diagnostic accuracy exceeded 95% in both entities. Eventually, asthma was categorized as a separate category within the respiratory system, and the diagnostic system accurately distinguished between uncomplicated asthma, cough variant asthma, and acute asthma exacerbation. In the table, computer predicted diagnoses are underlined, and physician diagnoses are presented in the first row at the top. Bold values indicate high accuracy of the computer prediction. AAE, acute asthma exacerbation; A. La., acute laryngitis; AURI, acute upper respiratory infection; CVA, cough variant asthma; Gast, gastrointestinal; L. Resp., lower respiratory; Neuro., neuropsychiatric; Resp., respiratory; Sin., sinusitis; Sys., systemic orgeneralized; U.A. uncomplicated asthma; U. Resp., upper respiratory.

as bacterial meningitis, for which accuracy was 0.93 (Supplementary Table 5).

To gain insight into how the diagnostic system generated a diagnosis prediction, we identified key clinical features driving the diagnosis prediction. For each feature, we identified which category of EHR clinical data it was derived from (for example, history of present illness, physical exam) and whether it was coded as a binary classification or categorical (Supplementary Table 7). The interpretability of the predictive impact of clinical features used in our diagnostic system allowed us to evaluate whether the prediction was based on clinically relevant features.

In terms of gastroenteritis, for example, the diagnostic system identified words such as 'abdominal pain' and 'vomiting' as key associated clinical features. The binary classifiers were coded such that the presence of a feature was denoted as '1' and absence was denoted as '0'. In this case, 'vomiting = 1' and 'abdominal pain = 1' were identified as key features for both chief complaint and history of present illness. Under physical examination, 'abdominal tenderness = 1' and 'rash = 1' were noted to be associated with this diagnosis. Interestingly, 'palpable mass = 0' was also associated, meaning that the patients predicted to have gastroenteritis usually did not have a palpable mass, which is consistent with human clinical experience. In addition to binary classifiers, there were also nominal categories in the schema. The feature of 'fever' with a text entry of greater than 39° C also emerged as an associated clinical feature driving the diagnosis of gastroenteritis. Laboratory and imaging features were not identified as strongly driving the prediction of this diagnosis, perhaps reflecting the fact that most cases of gastroenteritis are diagnosed without extensive ancillary tests.

We also compared the performance of diagnosis between our AI model and human physicians using 11,926 records from an independent cohort of pediatric patients. Twenty pediatricians in five groups with increasing levels of proficiency and years of clinical practice experience (see Methods section for description) manually graded 11,926 records. A physician in each group read a random subset of the raw clinical notes from this independent validation data set and assigned diagnoses. We evaluated the diagnostic performance of each physician group in each of the top 15 diagnosis categories using an F1 score (Table 2). Our model achieved an average F1 score higher than the two junior physician groups but lower than the three senior physician groups. This result suggests that this AI model may potentially assist junior physicians in diagnoses but may not necessarily outperform experienced physicians.

Here, we present an AI-based NLP model that can process free text from physicians' notes in ~~the EHR~~ to accurately predict the primary diagnosis in a pediatric ~~c~~population. The model was initially trained using a set of notes that were manually annotated by an expert team of physicians and informatics researchers. Once trained, the NLP information extraction model used deep learning techniques to automate the annotation process for notes from over 1.4 million patient encounters from a single institution in China. With the clinical features extracted and annotated by the deep NLP model, logistic regression classifiers were used to predict the primary diagnosis for each encounter. This system achieved excellent performance across all organ systems and subsystems, demonstrating a high level of accuracy for its predicted diagnoses compared with the initial diagnoses determined by an examining physician.

This diagnostic system demonstrated strong performance for two important categories of disease: common conditions that are frequently encountered in the population of interest, and dangerous or even potentially life-threatening conditions, such as acute asthma exacerbation and meningitis. Being able to predict common diagnoses as well as dangerous diagnoses is crucial for any diagnostic system to be clinically useful. For common conditions, there is a large pool of data to train the model, so we would expect a better performance with more training data. Accordingly, the performance

Table 2 | Illustration of diagnostic performance of our AI model and physicians

Disease conditions	Our model	Physicians				
		Physician group 1	Physician group 2	Physician group 3	Physician group 4	Physician group 5
Asthma	0.920	0.801	0.837	0.904	0.890	0.935
Encephalitis	0.837	0.947	0.961	0.950	0.959	0.965
Gastrointestinal disease	0.865	0.818	0.872	0.854	0.896	0.893
Group: 'Acute laryngitis'	0.786	0.808	0.730	0.879	0.940	0.943
Group: 'Pneumonia'	0.888	0.829	0.767	0.946	0.952	0.972
Group: 'Sinusitis'	0.932	0.839	0.797	0.896	0.873	0.870
Lower respiratory	0.803	0.803	0.815	0.910	0.903	0.935
Mouth-related diseases	0.897	0.818	0.872	0.854	0.896	0.893
Neuropsychiatric disease	0.895	0.925	0.963	0.960	0.962	0.906
Respiratory	0.935	0.808	0.769	0.89	0.907	0.917
Systemic or generalized	0.925	0.879	0.907	0.952	0.907	0.944
Upper respiratory	0.929	0.817	0.754	0.884	0.916	0.916
Root	0.889	0.843	0.863	0.908	0.903	0.912
Average F1 score	**0.885**	**0.841**	**0.839**	**0.907**	**0.915**	**0.923**

We used the F1score to evaluate the diagnosis performance across different groups (rows); our model, two junior physician groups (groups 1 and 2), and three senior physician groups (groups 3, 4, and 5) (see Methods section for description). We observed that our model performed better than junior physician groups but slightly worse than three experienced physician groups. Root is the first level of diagnosis classification.

of our system was especially strong for the common conditions of acute upper respiratory infection and sinusitis, both of which were diagnosed with an accuracy of 0.95 between the machine-predicted diagnosis and the human physician-generated diagnosis. In contrast, dangerous conditions tend to be less common and would have less training data. Despite this, a key goal for any diagnostic system is to achieve high accuracy for these dangerous conditions in order to promote patient safety. Our system was able to achieve this in several disease categories, as illustrated by its performance for acute asthma exacerbations (0.97), bacterial meningitis (0.93), and across multiple diagnoses related to systemic generalized conditions, such as varicella (0.93), influenza (0.94), mononucleosis (0.90), and roseola (0.93). These are all conditions that can have potentially serious and sometimes life-threatening sequelae, so accurate diagnosis is of utmost importance.

Another strength of this study was the massive volume of data that was used, with over 1.4 million records included in the analysis. It has been well documented that machine learning techniques improve as the amount of input data increases[10–12], so the large volume of encounters here contributed to the robustness of the diagnostic system. Furthermore, another strength was that the data inputs in this model were harmonized. This represents an improvement upon other techniques, such as mapping the attributes to a fixed format (Fast Healthcare Interoperability Resources), as was done recently in an AI-based analysis of EHR data[13]. Harmonized inputs describe the data in a consistent fashion and improve the quality of the data using machine learning capabilities[14]. These strengths of high volume of data, and harmonization of data inputs are key advantages of this model compared with other NLP frameworks that have been reported previously.

Our overall framework of automating the extraction of clinical data concepts and features to facilitate diagnostic prediction can potentially be applied across a wide array of clinical applications. In this study, we used primarily an anatomical or organ systems-based approach to the diagnostic classification. This broad generalized approach is often used in the formulation of differential diagnoses by physicians. Other strategies include using a pathophysiological or etiological approach (for example, 'infectious' versus 'inflammatory' versus 'traumatic' versus 'neoplastic'). The design of the diagnostic hierarchy decision tree can be adjusted to what is most appropriate for the clinical situation.

In terms of implementation, we foresee this type of AI-assisted diagnostic system being integrated into clinical practice in several ways. First, it could assist with triage procedures. For example, when patients come to the emergency department or to an urgent care setting, their vital signs, basic history, and notes from a physical examination by a nurse or midlevel provider could be entered into the framework, allowing the algorithm to generate a predicted diagnosis. These predicted diagnoses could help to prioritize which patients should get seen first by a physician. Some patients with relatively benign or non-urgent conditions may even be able to bypass the physician evaluation altogether and be referred for routine outpatient follow-up in lieu of urgent evaluation. This diagnostic prediction would help to ensure that physicians' time is dedicated to the patients with the highest and/or most urgent needs. By triaging patients more effectively, waiting times for emergency or urgent care may decrease, allowing improved access to care within a healthcare system of limited resources.

Another potential application of this framework is to assist physicians with the diagnosis of patients with complex or rare conditions. While formulating a differential diagnosis, physicians often draw upon their own experiences, and therefore the differential may be biased towards conditions that they have seen recently or that they have commonly encountered in the past. However, for patients presenting with complex or rare conditions, a physician may not have extensive experience with that particular condition. Misdiagnosis may be a distinct possibility in these cases. Using this AI-based diagnostic framework harnesses the power generated by data from millions of patients and would be less prone to the biases of individual physicians. In this way, a physician could use the AI-generated diagnosis to help to broaden his or her differential diagnosis and think of diagnostic possibilities that may not have been immediately obvious.

In conclusion, this study describes an AI framework to extract clinically relevant information from free text EHR notes to accurately predict a patient's diagnosis. Our NLP information model was able to perform the information extraction with high recall and precision across multiple categories of clinical data, and when

processed with logistic regression classifiers, was able to achieve high association between predicted diagnoses and initial diagnoses determined by a human physician. This type of framework may be useful for streamlining patient care, such as in triaging patients and differentiating between those patients who are likely to have a common cold from those who need urgent intervention for a more serious condition. Furthermore, as NLP processes become increasingly refined, these frameworks could become a diagnostic aid for physicians and assist in cases of diagnostic uncertainty or complexity, thus not only mimicking physician reasoning but augmenting it as well. Although this impact may be most obvious in areas in which there are few healthcare providers relative to the population, such as China, healthcare resources are in high demand worldwide, and the benefits of such a system are likely to be universal.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41591-018-0335-9.

Received: 18 July 2018; Accepted: 7 December 2018;
Published online: 11 February 2019

References

1. Hu, J., Perer, A. & Wang, F. *Data Driven Analytics for Personalized Healthcare*. (Springer Internatonal Publishing, Switzerland, Healthcare Information Management Systems: Cases, Strategies, and Solutions, 2016).
2. Nezhad, M.Z., Zhu, D.X., Sadati, N., Yang, K. & Levy, P. SUBIC: A supervised bi-clustering approach for precision medicine. *2017 16th Ieee International Conference on Machine Learning and Applications (Icmla)*. Preprint at https://arxiv.org/pdf/1709.09929.pdf (2017).
3. Hornberger, J. Electronic health records: a guide for clinicians and administrators. *JAMA* **301**, 110–110 (2009).
4. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
5. Kermany, D. S. et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122–1131 (2018).
6. Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410 (2016).
7. Erickson, B. J., Korfiatis, P., Akkus, Z. & Kline, T. L. Machine learning for medical imaging. *Radiographics* **37**, 505–515 (2017).
8. Wang, F., Zhang, P., Qian, B., Wang, X. & Davidson, I. Clinical risk prediction with multilinear sparse logistic regression. In *Proc. 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.145–154 (2014).
9. Turchin, A. et al. Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes. *J. Am. Med. Inform. Assoc.* **13**, 691–695 (2006).
10. Halevy, A., Norvig, P. & Pereira, F. The unreasonable effectiveness of data. *IEEE Intelligent Systems* **24**, 8–12 (2009).
11. Banko, M. & Brill, E. Scaling to very very large corpora for natural language disambiguation. In *Proc. 39th Annual Meeting Association for Computational Linguistics*. 26–33 (Association for Computational Linguistics, Stroudsburg, 2001).
12. Tsui, B. Y., et al. Creating a scalable deep learning based named entity recognition model for biomedical textual data by repurposing biosample free-text annotations. Preprint at https://www.biorxiv.org/content/biorxiv/early/2018/09/12/414136.full.pdf (2018).
13. Rajkomar, A. et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine* **1**, 18 (2018).
14. Wilkinson, M. D. et al. Comment: the fair guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).

Acknowledgements

This study was funded by the National Key Research and Development Program of China (2017YFC1104600 to H.L.), National Natural Science Foundation of China (81771629 to H.X. and 81700882 to J.X.), Guangzhou Women and Children's Medical Center, Guangzhou Regenerative Medicine and Health Guangdong Laboratory (Innovation and Startup Talents Program 2018GZR031001 to L.Z. and R.H.).

Author Contributions

H.L., B.T., H.N., W.C., S.L.B., G. Liu, D.S.K., X. S., C.C.S.V., P.T., H.S., J.C., L. H., J.Z., L.Z., R.H., S.H., G. Li, P.L., X.Z., Z.Z., L.P., H.C., R.L., S.L., Y.C., S.T., H.Y., X.H., W. He, W.L., Q.Z., J.J., W.Y., J.G., W.O., Y. Deng, Q.H., B. Wang, C.Y., Y.L., S.Z., Y. Duan, R.Z., S.G., C.L.Z., O.L., E.D.Z., G.K., X.W., C.W., N.N., J.X., W.X., B. Wang, W.W., J.L., B.P., C.B., D.X., W. He, S.H., Y.Z., W. Haw, M.G., A.T., C.-N.H., H.C., L.Z., H.X. and K.Z. collected and analyzed the data. X.H. and K.Z. conceived the project. K.Z., S.L.B., B.T., H.L., and H.X. wrote the manuscript. All authors discussed the results and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-018-0335-9.

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-018-0335-9.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to K.Z. or H.X.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Data collection. We conducted a retrospective study and obtained EHRs from 1,362,559 outpatient visits from 567,498 pediatric patients from the Guangzhou Women and Children's Medical Center, a major Chinese academic medical referral center. These records encompassed physician encounters for pediatric patients presenting to this institution from January 2016 to July 2017. The median age was 2.35 years (range 0 to 18 years, 95% confidence interval 0.2 to 9.7 years), and 40.11% were female (Supplementary Table 1). Disease prevalence from Supplementary Table 1 is derived from the official government statistics report from the Guangdong province[15]. All encounters included the primary diagnosis in the International Classification of Disease (ICD)-10 coding determined by the physician[16].The EHR system was developed by a Chinese vendor named Zesing Electronic Medical Records. A further 11,926 patient visit records from an independent cohort of pediatric patients from Zengcheng Women and Children's Hospital (Guangdong Province, China) were used for a comparison study between our AI system and human physicians.

The study was approved by the Guangzhou Women and Children's Medical Center and Zengcheng Women and Children's Hospital institutional review board and complied with the Declaration of Helsinki. Informed written consents were obtained from all participants at the initial hospital visit. Patient sensitive information was removed during the initial extraction of EHR data and EHR were de-identified. Data were stored in a fully HIPAA (Health Insurance Portability and Accountability Act)-compliant manner.

NLP model construction. We established a raw information extraction model, which extracted the key concepts and associated categories in EHR raw data and transformed them into reformatted clinical data in query–answer pairs (Extended Data 2). The reformatted chart grouped the relevant symptoms into categories, which increased interpretability by showing the exact features that the model relies on to make a diagnosis. Three physicians curated and validated the schemas, which encompassed chief complaint, history of present illness, physical examination, and laboratory reports. There were multiple components to the NLP framework: lexicon constructionl; tokenization; word embedding; schema construction; and sentence classification using long short-term memory (LSTM) architecture. The median number of records included in the training cohort for any given diagnosis was 1,677, but there was a wide range (4 to 321,948) depending on the specific diagnosis. Similarly, the median number of records in the test cohort for any given diagnosis was 822, but the number of records also varied (range of 3 to 161,136) depending on the diagnosis.

Lexicon construction. The lexicon was generated by manually reading sentences in the training data (approximately 1% of each class, consisting of over 11,967 sentences) and selecting clinically relevant words for the purpose of query–answer model construction. The keywords were curated by our physicians and were generated by using a Chinese medical dictionary[17], which is analogous to the unified medical language system (UMLS)[18] in the United States. Next, any errors in the lexicon were revised according to the physicians' clinical knowledge and experience, as well as expert consensus guidelines, based on conversations between two board-certified internal medicine physicians, one informatician, and one health information management professional. This procedure was iteratively conducted until no new concepts of history of present illness and physical examination were found. We then used these 11,967 sentences to train a word embedding model.

Schema design. The schema consists of a list of physician curated questions-and-answer pairs that the physician would use in extracting symptom information towards the diagnosis. Examples of questions are 'Does the patient have a fever?' and 'Is the patient coughing?'. The answer consists of a key_location and a numeric feature. The key_location encodes anatomical locations such as lung or gastrointestinal tract. Therefore, the value is either a categorical variable or a binary number depending on the feature type. Then, we constructed a schema for each type of medical record data: the history of present illness and chief complaint, physical examination, laboratory tests. We then applied this schema towards the text re-formatting model construction.

The rationale for this schema design was to maximize data interoperability across hospitals for future study. The pre-defined space of query–answers pairs simplifies the data interpolation process across EHR systems from multiple hospitals. Also, providing clinical information in reduced formats can help protect patient privacy compared to providing raw clinical notes that could be patient-identifiable. Even with removal of patient-identifiable variables, the style of writing in the EHR may potentially reveal the identity of the examining physician, as suggested by advances in stylometry tools[19], which could increase patient identifiability.

Tokenization and word embedding. Due to the lack of publicly available community annotated resources for the clinical domain in Chinese, we built standard data sets for word segmentation. The tool used for tokenization was Mecab (https://github.com/taku910/mecab), with our curated lexicons as the optional parameter. We had a total of 4,363 tokens. We used word2vec from the python Tensorflow package (1.9.0) to embed the 4,363 tokens with 100 features, to represent the semantics and similarities of the word in the high dimensional space.

LSTM model training set and test set construction. We curated a small data set for training the query–answer extraction model. We manually annotated the query–answer pairs in our training ($n = 3,564$) and validation ($n = 2,619$) cohort. For questions with binary answers, we used 0,1 to indicate whether the text gives a no or yes. For example, given the text snippet 'the patient has a fever', the query 'Does the patient have a fever?' is assigned a value of 1. For queries with categorical or numerical values, we assigned each a pre-defined categorical answer.

Our free text harmonization process was modeled using the attention-based LSTM described previously[20]. We implemented the model using Tensorflow and trained the model with 200,000 steps. We applied our NLP model to all EHR physician notes and converted them into a structured format, in whicheach record contained data in query–answer pairs (Extended Data 2). We did not tune the hyperparameters but relied on either default or commonly used settings of hyperparameters for the LSTM model. We used a default of 128 hidden units per layer as reported in multiple publications[21,22] and two layers of LSTM cells as suggested by the commonly adopted bidirectional LSTM;[23] we used a default learning rate of 0.001 from Tensorflow.

Hierarchical multi-label diagnosis model. *Diagnosis hierarchy curation.* The diagnosis hierarchy was curated by at least two US board-certified physicians and two Chinese board-certified physicians. An anatomically based classification system was used for the diagnostic hierarchy, as this is a common practice for formulating a differential diagnosis when a human physician evaluates a patient. First, the diagnoses were separated into general organ systems (for example, respiratory, neuropsychiatric, or gastrointestinal). Within each organ system, there was a subdivision into subsystems (for example, upper respiratory and lower respiratory). A separate category was labeled 'systemic or generalized' in order to include conditions that affected more than one organ system and/or were more general in nature (for example, mononucleosis or influenza).

Model training and validation process. The data from the query–answer model consist of a mix of categorical variables and yes or no binary answers. Therefore, we used the hot-one encoding scheme to first convert both the categorical and binary answers to a unified binary feature by visit matrix. The data were then randomly split into a training cohort, consisting of 70% of the total visit records, and a test cohort, comprising the remaining 30%. We then annotated each visit record in the training and test cohort by constructing a query–answer membership matrix.

For each intermediate node, we trained a multiclass linear logistic regression classifier based on the immediate child terms. All the subclasses of the child terms were collapsed to the level of the child terms. The one versus rest multiclass classifier was trained using Sklearn class LogisticRegression with the default l1 regularization penalty (Lasso), simulating situations in which physicians rely on a limited number of symptoms to make a diagnosis. The inputs were in query–answer pairs as described above. To further evaluate the model, we also generated the receiver operating characteristic—area under curves (ROC-AUC) (Supplementary Table 8) to evaluate the sensitivity and specificity of our multiclass linear logistic regression classifiers. We also examined the robustness of our classification models using fivefold cross-validation (Supplementary Table 9).

Hierarchical clustering of disease. We correlated the mean profile of the feature membership matrix using the Pearson correlation. Hierarchical clustering was carried out using the *clustermap* function of the Python Seaborn package with default parameters.

To evaluate the robustness of the clustering result (Extended Data 1), we first randomly split the data in half, with one half for training and the other for testing, and regenerated the two cluster maps for the training and test data independently. We assigned the leaves in both the training and test cluster maps to ten classes by cutting the associated dendrogram at the corresponding height independently. The class assignment concordance between the training and test data was evaluated using the adjusted Rand index (ARI)[24]. An ARI value closer to 1 indicates higher concordance between training class assignment and test class assignment, whereas an ARI closer to 0 indicates close to the null background. We observed a high ARI of 0.8986 between the training and test class assignments, suggesting that our cluster map is robust. In several instances, the system clustered diagnoses with related ICD-10 codes, illustrating that it was able to detect trends in clinical features that align with a human-defined classification system. However, in other instances, it clustered together related diagnoses but did not include other very similar diagnoses within this cluster. For example, it grouped 'asthma' and 'cough variant asthma' into the same cluster, but did not include 'acute asthma exacerbation', which was instead grouped with 'acute sinusitis'. Several similar pneumonia-related diagnosis codes were also spread across several different clusters instead of being grouped together. In many instances, it successfully established broad grouping of related diagnoses even without any directed labeling or classification system in place, suggesting that the clinical features that we developed capture the key similarities and differences between the conditions that we intend to model and diagnose.

Comparison of performance of our AI system with that of human physicians. We conducted a study to compare the performance of our AI system with that of human physicians using 11,926 records from an independent cohort of pediatric patients from Zengcheng Women and Children's Hospital, Guangdong Province, China. We chose 20 pediatricians in five groups with increasing levels of proficiency and years of clinical practice experience (four in each level) to manually grade 11,926 records. These five groups are: senior resident physicians with more than 3 years' practice experience, junior physicians with 8 years' practice experience, midlevel physicians with 15 years' practice experience, attending physicians with 20 years' practice experience, senior attending physicians with more than 25 years' practice experience. A physician in each group read a random subset of 2,981 clinical notes from this independent validation dataset and assigned a diagnosis. Each patient record was randomly assigned and graded by four physicians (one in each physician group). We evaluated the diagnostic performance of each physician group in each of top 15 diagnosis categories using an F1 score (Table 2).

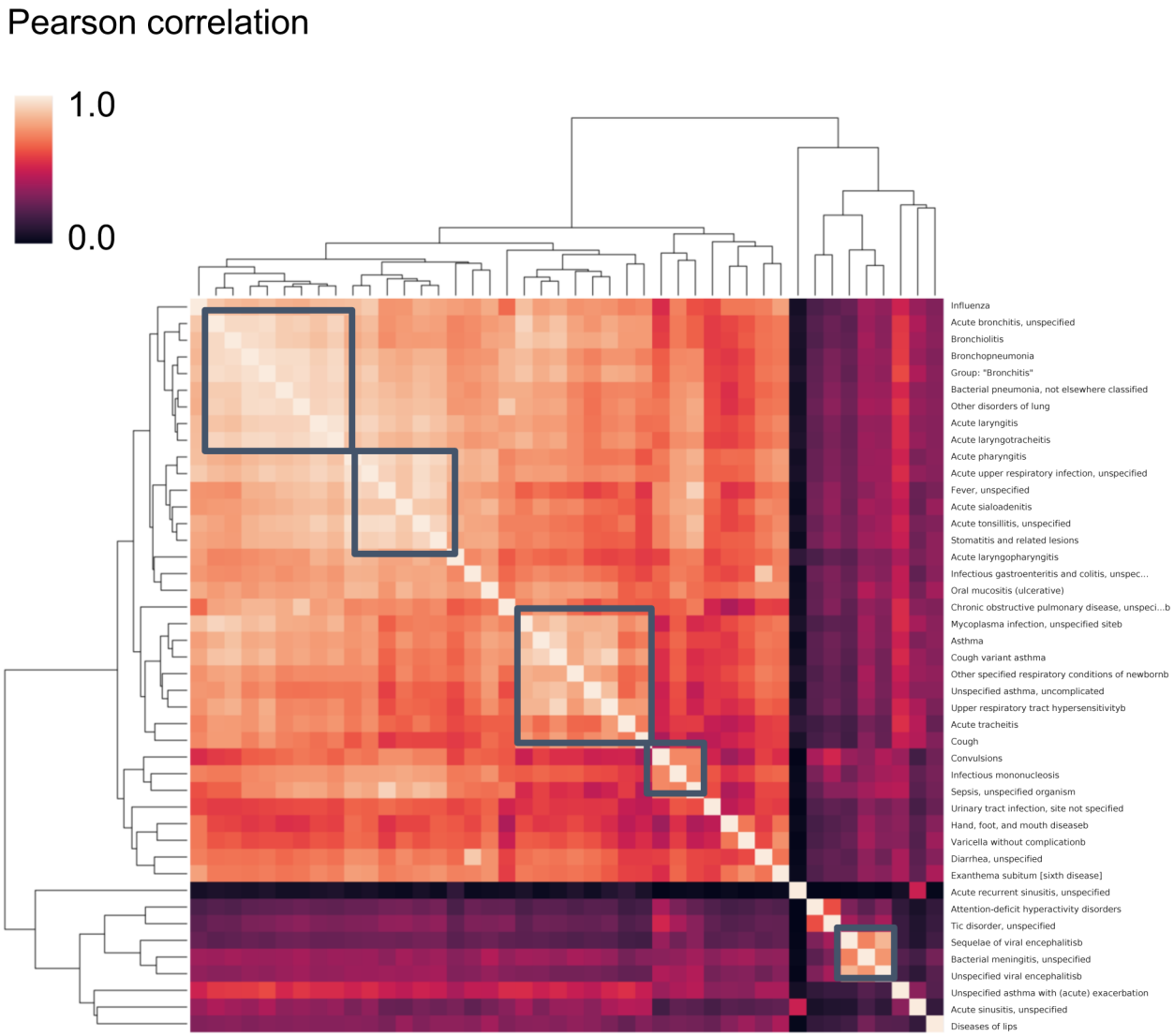
Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

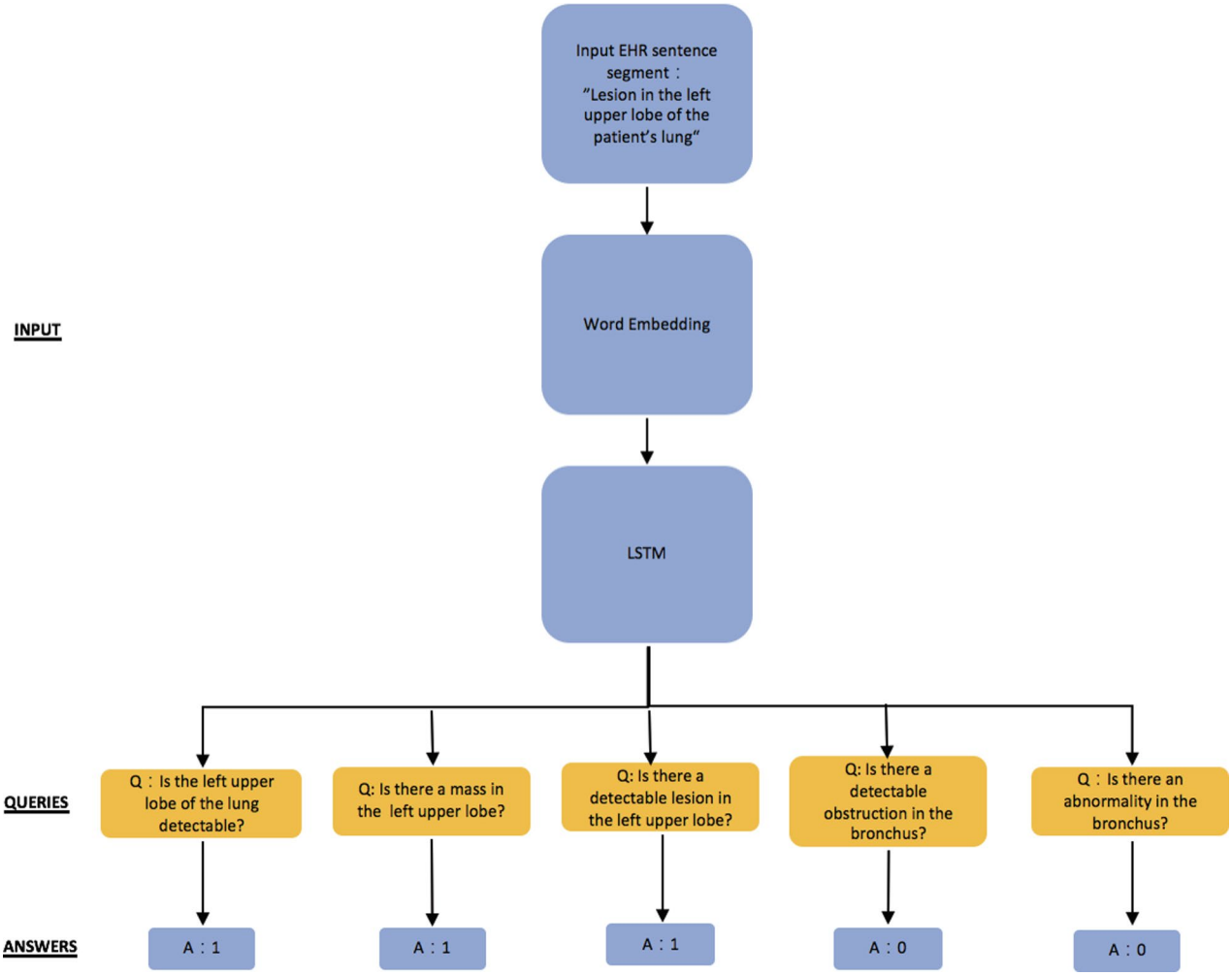
We have made available the Jupyter notebook that we used in constructing and validating the hierarchical logistic regression models: https://s3.cn-north-1.amazonaws.com.cn/ped.emr/Data/hierachical_logistic_regression.ipynb. To protect patient confidentiality, we have deposited de-identified aggregated patient data in a secured and patient confidentiality compliant cloud in China in concordance with data security regulations. Data access can be requested by writing to the corresponding authors. All data access requests will be reviewed and (if successful) granted by the Data Access Committee.

References

15. Liang, Y., Chen, Z., Huang, X. & Zeng, L. Analysis of the disease spectrum of hospitalized children in guangdong province. *Chin. Med. J. (Engl)* **1**, 414–418 (2013).
16. WHO. *International Statistical Classification of Diseases and Related Health Problems.* (World Health Organization, 2004).
17. *English–Chinese Medical Dictionary* (英汉医学大词典) (Shanghai Scientific and Technical Publishers (上海科学技术出版社), 2015).
18. Lindberg, D. A. B., Humphreys, B. L. & Mccray, A. T. The unified medical language system. *Methods Inf. Med.* **32**, 281–291 (1993).
19. Tweedie, F. J., Singh, S. & Holmes, D. I. Neural network applications in stylometry: the federalist papers. *Computers and the Humanities* **30**, 1–10 (1996).
20. Luong, M.-T., Pham, H. & Manning, C. D. Effective approaches to attention-based neural machine translation. Preprint at https://arxiv.org/abs/1508.04025 (2015).
21. Lipton, Z.C., Kale, D.C. & Wetzel, R.C. Phenotyping of clinical time series with LSTM recurrent neural networks. Preprint at https://arxiv.org/pdf/1510.07641.pdf (2015).
22. Peng, X.B., Andrychowicz, M., Zaremba, W. & Abbeel, P. Sim-to-real transfer of robotic control with dynamics randomization. *IEEE International Conference on Robotics and Automation (ICRA)* 3803–3810 (2018).
23. Graves, A. & Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* **18**, 602–610 (2005).
24. Yeung, K.Y. & Ruzzo, W.L. Details of the adjusted rand index and clustering algorithms supplement to the paper 'an empirical study on Principal Component Analysis for clustering gene expression data. Available at http://faculty.washington.edu/kayee/pca/supp.pdf (2011).



Extended Data 1 | Unsupervised clustering of NLP extracted textual features from pediatric diseases. The diagnostic system analyzed the EHRs in the absence of a defined classification system. This grouping structure reflects the detection of trends in clinical features without pre-defined labeling or human input. The clustered blocks are marked with the boxes with grey lines.



Extended Data 2 | Design of the natural language processing information extraction model. Segmented sentences from the raw text of the EHR were embedded using word2vec. The LSTM model then generated the structured records in a query–answer format. This schematic illustrates the process using the free-text 'lesion in the upper left lobe of patient's lung' as an example.

nature research

Corresponding author(s): Kang Zhang

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☒	☐	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	The electronic health records used in this study was developed by a Chinese vendor named Zesing Electronic Medical Records.
Data analysis	Mecab (URL: https://github.com/taku910/mecab) was used for tokenization. word2vec from the python Tensorflow package was used to embed the 4363 tokens with 100 features, to represent the semantics and similarities of the word in the high dimensional space. Python seaborn package was used to generate hierarchical clustering.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

We have made available the Jupyter notebook that we used in constructing and validating the hierarchical logistic regression models: https://s3.cn-

north-1.amazonaws.com.cn/ped.emr/Data/hierachical_logistic_regression.ipynb. To protect patient confidentiality, we have deposited de-identified aggregated patient data in a secured and patient confidentiality compliant cloud in China in concordance with data security regulations. Data access can be requested by writing to the corresponding authors. All data access requests will be reviewed and (if successful) granted by the Data Access Committee

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Electronic health records were collected from 1,362,559 outpatient patient visits from the Guangzhou Women and Children's Medical Center. (See Method). It has been well documented that machine learning techniques improve with a greater amount of input data, so the large volume of encounters here contributed to the robustness of the diagnostic system (See Discussion).
Data exclusions	We did not exclude any data
Replication	No experimental replication was attempted.
Randomization	The data was split into a training cohort, consisting of 70% of the total visit records, and a testing cohort, comprised of the remaining 30%. (See Method)
Blinding	Blinding is not applicable since only electronic health records, in which patient sensitive information was removed during the initial extraction of EHR data and EHR were de-identified, were used to evaluate performance of AI system v.s. human physicians.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Electronic health records of 1,362,559 outpatient visits from 567,498 pediatric patients from the Guangzhou Women and Children's Medical Center were collected. These records encompassed physician encounters for pediatric patients presenting to this institution from January 2016 to July 2017. The median age was 2.35 years (range: 0 to 18, 95% confidence interval: 0.2 to 9.7 years old), and 40.11% were female. 11,926 patient visit records from an independent cohort of pediatric patients from Zhengcheng Women and Children's Hospital (Guangdong Province, China) were used for a comparison study between our AI system and human physicians.
Recruitment	Electronic health records of 1,362,559 outpatient visits from 567,498 pediatric patients and 11,926 patient visit records were collected from Guangzhou Women and Children's Medical Center and Zhengcheng Women and Children's Hospital. The study was approved by the Guangzhou Women and Children's Medical Center and Zhengcheng Women and Children's Hospital institutional review board and ethics committee and complied with the Declaration of Helsinki. Consents were obtained from all participants at the initial hospital visit. Patient sensitive information was removed during the initial extraction of EHR data and EHR were de-identified. A data use agreement was composed and upheld by all institutions involved in the data collection and analysis. Data were stored in a fully HIPAA-compliant manner.