JAMIA Open, 3(4), 2020, 567–575
doi: 10.1093/jamiaopen/ooaa063
Advance Access Publication Date: 5 December 2020
Research and Applications

Research and Applications

Machine learning identifies girls with central precocious puberty based on multisource data

Liyan Pan[1,†], Guangjian Liu[1,†], Xiaojian Mao[2] and Huiying Liang[1]

[1]Institute of Pediatrics, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, China and [2]Department of Genetics and Endocrinology, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, China

*These authors contributed equally to this work.
Corresponding Author: Huiying Liang, PhD, Institute of Pediatrics, Guangzhou Women and Children's Medical Center, 9 Jinsui Road, Tianhe District, Guangzhou 510623, Guangdong, China (lianghuiying@hotmail.com)

Received 20 August 2020; Revised 26 October 2020; Editorial Decision 29 October 2020; Accepted 17 November 2020

ABSTRACT

Objective: The study aimed to develop simplified diagnostic models for identifying girls with central precocious puberty (CPP), without the expensive and cumbersome gonadotropin-releasing hormone (GnRH) stimulation test, which is the gold standard for CPP diagnosis.

Materials and methods: Female patients who had secondary sexual characteristics before 8 years old and had taken a GnRH analog (GnRHa) stimulation test at a medical center in Guangzhou, China were enrolled. Data from clinical visiting, laboratory tests, and medical image examinations were collected. We first extracted features from unstructured data such as clinical reports and medical images. Then, models based on each single-source data or multisource data were developed with Extreme Gradient Boosting (XGBoost) classifier to classify patients as CPP or non-CPP.

Results: The best performance achieved an area under the curve (AUC) of 0.88 and Youden index of 0.64 in the model based on multisource data. The performance of single-source models based on data from basal laboratory tests and the feature importance of each variable showed that the basal hormone test had the highest diagnostic value for a CPP diagnosis.

Conclusion: We developed three simplified models that use easily accessed clinical data before the GnRH stimulation test to identify girls who are at high risk of CPP. These models are tailored to the needs of patients in different clinical settings. Machine learning technologies and multisource data fusion can help to make a better diagnosis than traditional methods.

Key words: central precocious puberty; GnRH stimulation test; machine learning; multisource data

INTRODUCTION

Central precocious puberty (CPP) is a disease caused by premature activation of the hypothalamic-pituitary-gonadal (HPG) axis with clinical pubertal symptoms among girls under 8 years old and boys under 9 years of age. With the continuous improvement of living standards and the aggravation of environmental pollution, children are more exposed to endocrine disruptors, causing an increase in the incidence and prevalence of CPP.[1,2] For example, in Korea, the overall incidence among girls increased 4.7 times from 89.4 to 415.3 per 100 000 during the period from 2008 to 2014.[3] CPP has the potential to compromise adult height and even cause social psychological disturbances. Moreover, girls with CPP have an increased risk of breast or cervical cancer.[4] Therefore, early diagnosis and treatment are essential to girls with improper secondary sexual development.

LAY SUMMARY

- The basal hormone test had the highest diagnostic value for a Central Precocious Puberty (CPP) diagnosis.
- Image features extracted directly from bone age X-ray images perform better in diagnosis than the subjective bone age value from examination reports.
- Multisource and heterogeneous data fusion could increase model performance in medical diagnosis.
- Three different models with good performance were selected to simplify the CPP diagnosis flow and were tailored to the needs of different clinical settings.
- Machine learning could assist in CPP diagnosis without cumbersome laboratory tests.

Besides CPP, there is another type of precocious puberty (PP) called peripheral PP or non-CPP, which has similar clinical characteristics such as breast or uterus development with CPP, but without HPG axis activation. Non-CPP has a low risk of height compromise and its pubertal symptoms can subside spontaneously. It is difficult to distinguish between the two diseases, unless using the gold standard for CPP, that is, the gonadotropin-releasing hormone (GnRH) or GnRH analog (GnRHa) stimulation test. This is an expensive and time-consuming test that requires multiple blood samples during the whole process resulting in much patient suffering. In nontertiary or community hospitals with limited resources, this cumbersome test is not available. Hence, depending on access to the GnRH stimulation test, the diagnosis and treatment of CPP may be delayed. Several previous studies have tried to determine one adequate blood sampling time to simplify the stimulation test.[5–7] The adequate sampling time identified across studies differed greatly. Other studies targeted investigating factors such as basal sex hormone levels,[8] pelvic ultrasound,[9–11] or dental maturity[12] that may relate to CPP diagnosis. However, the cutoff values of these factors ranged widely.

Our previous study utilized data from 1757 girls who were diagnosed with precocious puberty and had undergone the GnRHa stimulation test before the age of 9 years.[13] We built models based on machine learning algorithms to identify patients with CPP. The results showed the significance of basal serum luteinizing hormone (LH), follicle-stimulation hormone (FSH), and insulin-like growth factor 1 (IGF-1) in differentiating CPP and non-CPP. Using a combination of basal hormone tests, secondary sexual characteristics, and data from examination reports, the model achieved a sensitivity of 77.94%, a specificity of 87.66% and an area under the curve (AUC) of 0.90 based on 436 patients.

Studies from other fields have indicated that multisource data may perform better than single-source data.[14–16] In the multimedia field, features for one modality (eg, video) can be better learned by the models if multiple modalities (eg, audio, video, and text) being learned together.[17] In medical image analysis, different views of the same part can provide comprehensive information to make a decision. In clinics, patients usually need to take laboratory tests or medical image examinations, in addition to physical examinations. These similar approaches across fields prompted us to consider fusing patient data from different sources (clinical, laboratory, examination, etc.) to make a better diagnosis of CPP. Recently, machine learning and deep learning technologies have been widely used in diagnostic models and medical image analysis. They support information fusion based on multisource data.

In this article, our first objective was to explore the diagnostic value of data from different sources for CPP diagnosis. Then, we investigated whether adding data from other sources would improve the performance of the prediction models. Finally, models with good performance were selected to simplify the CPP diagnosis flow and were tailored to the needs of different clinical settings.

METHODS

Participants

In this study, girls with secondary sexual characteristics onset under the age of 8 were enrolled from the Pediatric Endocrinology Department of Guangzhou Women and Children's Medical Center. Individuals with genetic disorders, tumors, lesions, McCune-Albright syndrome, neurofibromatosis, ovarian cysts, or other diseases and those taking hormone medications were excluded from this study. All patients underwent the GnRHa stimulation test. Girls meeting the following eligibility criteria were diagnosed with CPP: (1) peak LH concentration $\geq$ 10 IU/L or peak LH concentration $\geq$ 5 IU/L combined with a ratio of peak LH to FSH value $\geq$ 0.6 and (2) onset of secondary sexual characteristics under the age of 8 years. Girls who did not fulfill all the above-mentioned criteria were diagnosed with GnRH-independent precocious puberty, which was referred as non-CPP in this study. These diagnostic criteria for CPP assessment are widely used in China[18] and many other countries.[19] In the current traditional CPP diagnosis pathway (Figure 1 and Supplementary Material), patients with secondary sexual development before 8 years old first undergo a physical examination, which is recorded in electronic medical records (EMRs). Then, laboratory testing (LAB), bone age (BA) X-ray imaging, and pelvic ultrasonography (US) are suggested. After all these tests and examinations, the GnRHa stimulation test is used to make a gold standard diagnosis. In this study, the data obtained using these different approaches from the same patient were considered as data from different sources. We had in total of four data sources here: EMR, LAB, BA, and US.

Ethics

This study was approved by the institutional review board of Guangzhou Women and Children's Medical Center and conducted in accordance with the ethical guidelines of the Declaration of Helsinki of the World Medical Association. The requirement to obtain informed consent was waived because of the retrospective nature of the study. The data used in this study were anonymous, and no identifiable personal data of the patients were available for the analysis.

Statistical analysis

The population characteristics are presented using the mean and standard deviation (SD) for quantitative data and number (%) for categorical data. Comparisons between two groups were performed

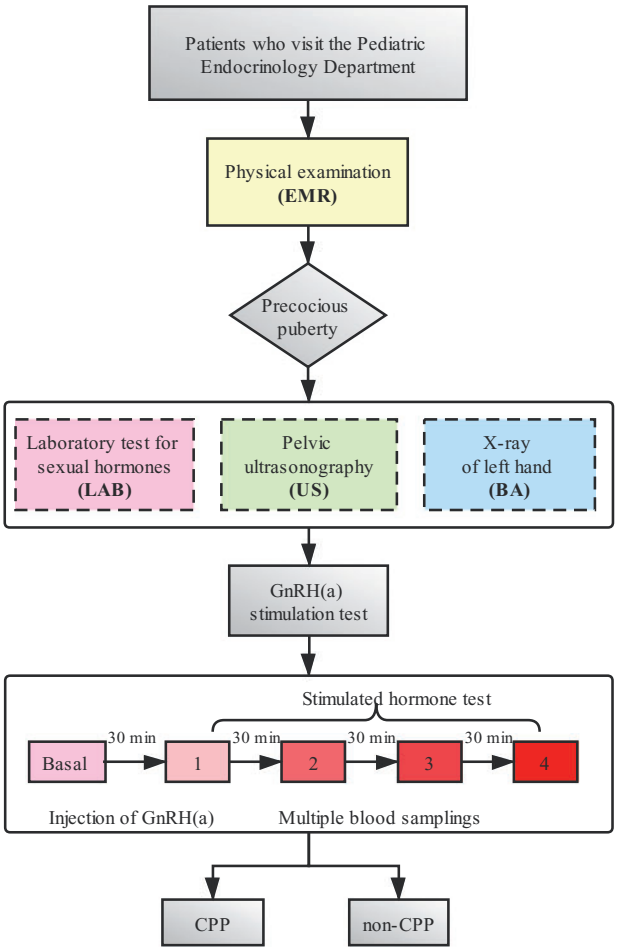

Figure 1. Current overall diagnosis flow for CPP. Different colors represent different data sources. EMR: electronic medical record. LAB: laboratory. US: ultrasound. BA: bone age.

using chi-square or Fisher's exact tests. The statistical analysis was performed using SPSS 22.0.

Data preprocessing

BA X-ray imaging is important in defining the advancement of bone maturity, which is directly associated with puberty. Here, we learned from BA value prediction through a deep learning method based on X-ray images. Features that are used to predict BA value have the potential to represent skeletal maturity. Thus, high-dimensional imaging data can be transformed into extracted 117 dimensional features using the DeepTW3 model.[20] In this study, features from the BA source existed in two forms: BA advancement from examination reports and image features extracted from X-ray images. Here, we used BA ratio and BA images to represent the features from BA reports and BA images, respectively.

For unstructured data such as medical records from the EMR source and examination reports from the BA and US sources, variables and their corresponding values were firstly extracted through regex matching. Then a manual checkout was performed to verify its effectiveness. Extracted features mainly contain information associated with secondary sexual development, such as ovarian volumes and Tanner stage for breast and pubic hair. We eliminated data with a missing rate over 60%. The missing rates of remaining data range from 0.62% to 50.12%. Missing values for continuous variables were filled with mean values of samples in the corresponding age group. For discrete variables such as development degree, missing values were empirically imputed with the least stage.

Model development

To investigate the diagnostic value of each single-source data and to validate whether multisource, heterogeneous data can improve model performance in CPP diagnosis, we designed models using each single-source data and different data combinations. Here, we used feature-level data fusion, which concatenated features from different sources directly. As tree-based ensemble classifiers in machine learning are easy to interpret, we used Extreme Gradient Boosting (XGBoost) classifier[21] as the basic classifier to develop CPP diagnostic models. XGBoost is a tree boosting and effective algorithm. It works by first minimizing errors of existing decision tree models and then training a sequence of models.

To evaluate the performance of all the models in the same test set, we randomly selected 20% of the patients who had data from all four sources as the independent test set, and the remaining patients were used as the training set to train different models. In the training stage, an inner k-fold crossvalidation ($k = 10$) was used for parameter tuning and model construction. In detail, the training set was randomly partitioned into 10 subsets of equal size, nine for training and the other one for validation. Grid search was used to tune the model hyperparameters. Sensitivity, specificity, AUC, and Youden index were used to assess model performance on the test set. Youden index is an index defined as the overall correct classification rate at the optimal cutoff point minus one, [sensitivity + specificity − 1]. The sensitivity and specificity computed here were rates with the optimal cutoff point based on the receiver operating characteristics (ROC) curve. Comparisons between models with the same AUC and Youden index were considered. The overall flow chart of the model development is shown in Figure 2. It included three main procedures: data preprocessing, data fusion, and model construction.

In the gradient boosting model XGBoost, feature importance is given by the selection frequency as a tree node. In the 10-fold cross-validation, each feature had feature importance each time, and the final feature importance for each one was computed as an average of all the values. In the model trained with the comprehensive data, feature importance for each variable was analyzed. All modeling analyses were performed in Python 3.5 using packages such as Scikit-learn, Pandas and Numpy.

RESULTS

Characteristics

From a total of 2523 patients who had undergone the GnRHa stimulation test, 1153 patients were diagnosed with CPP and 1370 with non-CPP. The numbers of patients in each group based on data source were 2523 (LAB), 1655 (EMR), 1612 (BA), and 2007 (US). Among them, 900 patients had all the data from the four heterogeneous data sources. The characteristics of patients grouped by data sources are presented in Table 1.

The mean ages of the CPP and non-CPP girls were 7.05 ± 1.13 years and 7.47 ± 1.09 years, respectively, which were significantly different ($P < .001$). With the exception of prolactin, the concentrations of the other laboratory variables, such as basal LH, FSH, and IGF-1, were significantly higher in the CPP group than those in the non-CPP group ($P < .001$). The girls diagnosed with CPP had higher

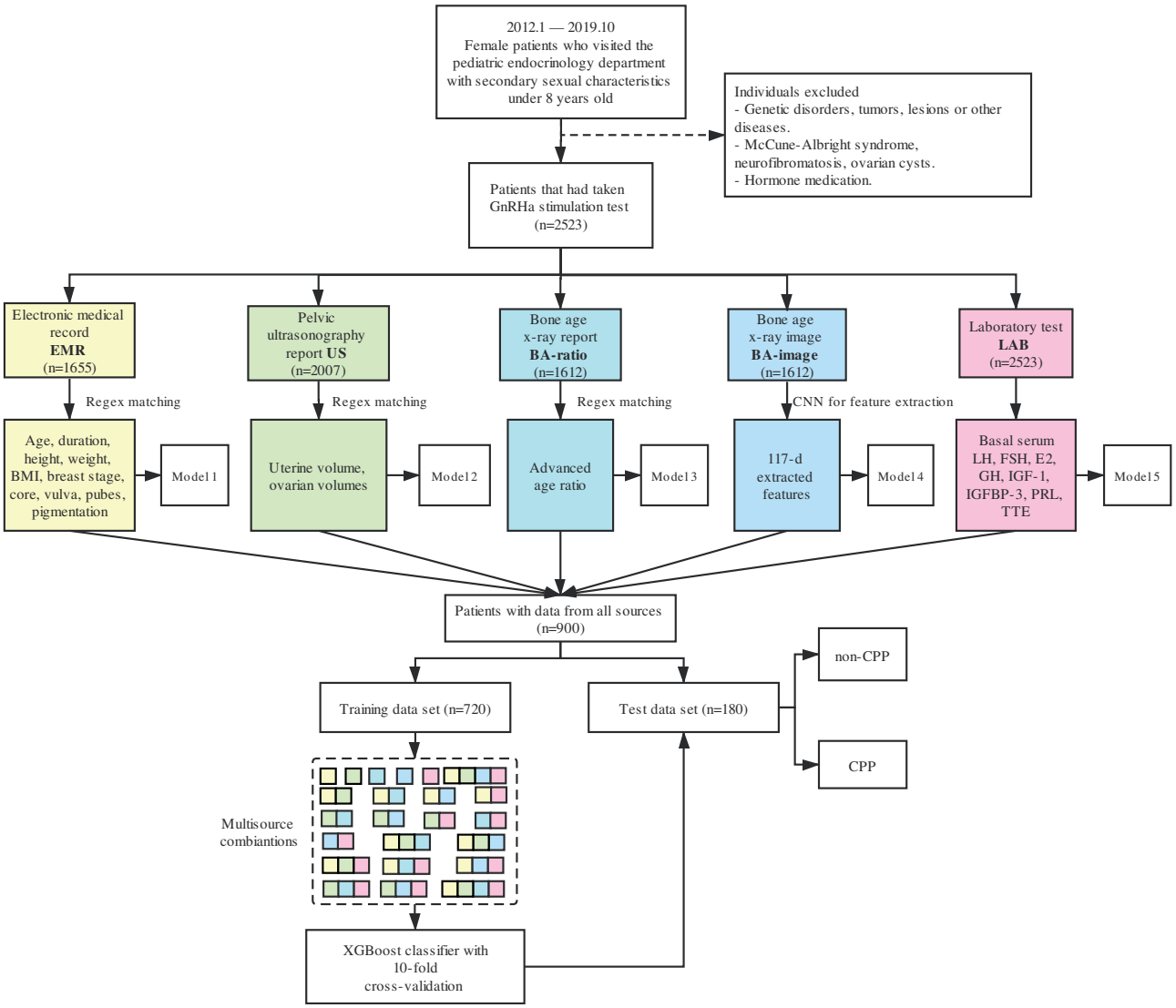

Figure 2. Flow chart for CPP diagnostic model development. Different colors represent different data sources.

height, weight, and BMI and more advanced breast development than those with non-CPP. Both BA value and age ratio (bone age advancement degree, the ratio of bone age to chronological age) were significantly more advanced in the CPP group ($P < .001$). All the variables extracted from the pelvic ultrasonography report showed that the ovary and uterine development of CPP girls were more advanced than those of non-CPP girls.

Model performance

We developed a total of 23 models with different data source combinations and validated all the models on the same independent test set consisting of 180 patients with comprehensive data. Sensitivity, specificity, AUC, and Youden index were computed for each model. The performance of five models based on single-source data and three models based on multiple data sources are presented in Table 2. The performances of all the models are shown in Figure 3. Detailed information about all the performances and relative comparisons are listed in Supplementary Table S2 (see Supplementary Material). Among the models using single-source data, the model based on LAB data achieved the best performance with an AUC of 0.85 and Youden index of 0.61 on the test data set. The model trained merely by BA data (BA ratio or BA image) achieved high sensitivity but low specificity. Performance increased for most of the models when LAB data were added. The model using data from the LAB and EMR sources achieved an AUC of 0.86 and Youden index of 0.65. Models based on data combinations without LAB showed worse performance than the models based on only LAB data or data combinations including LAB. As expected, among all the models, the model trained with all four data sources performed best, with an AUC of 0.88 and Youden index of 0.64.

Learning curves

Since the sample size in each single-source model's training set ranged from 720 to 2343, to investigate whether the performance would be affected by sample size, we plotted learning curves for each single-source dataset (Figure 4). The horizontal axis represents the sample size, and the vertical axis represents the model performance that varied with the sample size. The black dots represent the performance achieved at the maximum size of the training set with comprehensive data ($n = 720$). From the learning curves, as the sam-

Table 1. Characteristics of patients enrolled in this study

Characteristics	Non-CPP	CPP	*P*-value
Laboratory parameters ($n = 2523$)	1370	1153	
LH (IU/L)	0.11 (0.17)	0.78 (1.12)	<.001
FSH (IU/L)	1.83 (1.23)	2.94 (1.64)	<.001
GH (ng/mL)	3.18 (2.96)	4.13 (4.01)	<.001
IGF-1 (ng/mL)	232.25 (65.73)	295.19 (90.74)	<.001
IGFBP-3 (μg/mL)	4.69 (0.69)	4.85 (0.64)	<.001
E2 (pmol/L)	106.03 (59.17)	120.79 (56.03)	<.001
PRL (ng/mL)	9.42 (7.05)	8.765 (5.02)	.007
TTE (nmol/L)	0.80 (0.38)	0.90 (0.47)	<.001
Clinical parameters ($n = 1655$)	881	794	
Age (years)	7.05 (1.13)	7.47 (1.09)	<.001
Duration (months)	8.15 (10.86)	9.95 (10.44)	<.001
Height (cm)	126.11 (8.60)	129.95 (8.83)	<.001
Weight (kg)	26.04 (5.35)	28.33 (5.38)	<.001
BMI (kg/m^2)	16.22 (2.01)	16.66 (1.96)	<.001
Breast, tanner stage			<.001
1	72 (60.50)	47 (39.50)	
2	418 (67.97)	197 (32.03)	
3	347 (46.33)	402 (53.67)	
4	43 (23.12)	143 (76.88)	
5	1 (16.67)	5 (83.33)	
Core			.027
Yes	724 (51.42)	684 (48.58)	
No	157 (58.80)	110 (41.20)	
Vulva, tanner stage			.009
1	833 (53.57)	722 (46.43)	
2	44 (41.90)	61 (58.10)	
3	4 (26.67)	11 (73.33)	
Pubes, tanner stage			.002
1	837 (51.99)	773 (48.01)	
2	43 (71.67)	17 (28.33)	
3	1 (20.00)	4 (80.00)	
Pigmentation			.137
Yes	57 (60.00)	38 (40.00)	
No	824 (52.15)	756 (47.85)	
Bone age information ($n = 1612$)	897	715	
BA value (years)	8.70 (1.61)	9.72 (1.49)	<.001
Age ratio	1.25 (0.20)	1.32 (0.29)	<.001
Pelvic ultrasonography ($n = 2007$)	1048	923	
Left ovarian volume (mL)	2.04 (1.34)	2.46 (1.56)	<.001
Right ovarian volume (mL)	1.94 (1.32)	2.26 (1.17)	<.001
Uterine volume (mL)	1.57 (1.33)	2.69 (1.90)	<.001

LH: luteinizing hormone; IGF-1: insulin-like growth factor-1; FSH: follicle-stimulation hormone; PRL: prolactin; GH: growth hormone; E2: estradiol; BMI: body mass index; TTE: testosterone; BA: bone age.

ple size in the training set became larger and extended to its maximum size, higher AUC and Youden index values were obtained. It is also clear that the sample size for the multisource models ($n = 720$) is large enough to achieve a pretty good performance, although not the best performance.

Feature importance

To assess each variable's contribution to the model that achieved the best performance based on multi-source data (LAB+EMR+BA image+US), feature importance was computed for all 138 variables. The importance of the top 20 variables is plotted in Figure 5. Basal LH contributed the most to the prediction, followed by uterus volume and height. Age, weight, basal FSH, and IGF-1 from the EMR and LAB sources also had relatively high importance. In addition to the above variables, the variables with the prefix "Img-" shown in Figure 5 were features extracted from BA images, which represented the maturity of the epiphysis.

DISCUSSION

In this study, we built models based on different data source combinations to determine appropriate strategies for CPP diagnosis. The best performance was achieved with the model based on comprehensive data of four sources including LAB, EMR, BA image, and US, with an AUC of 0.88 and Youden index of 0.64. A significant strength of our study is the utilization of all the heterogeneous information before the stimulation test, including data from medical reports and images, to construct the models. Our findings indicated that models using machine learning and deep learning algorithms for CPP diagnosis reveal a trend that may be able to replace the

Table 2. Model performance of models using different data source combinations

Data sources					Sensitivity	Specificity	AUC	Youden index
LAB	EMR	BA		US				
		BA ratio	BA image					
●					76.62	84.62	0.85	0.61
	●				53.25	79.81	0.72	0.33
		●			81.82	38.46	0.62	0.20
			●		88.31	47.12	0.71	0.35
				●	66.23	75.96	0.76	0.42
	●		●	●	67.53	82.69	0.80	0.50
●	●				81.82	82.69	0.86	0.65
●	●		●	●	85.71	77.88	0.88	0.64

The first five lines show performance of models with single-source data. The last three lines show performance of selected models.
LAB: laboratory; EMR: electronic medical records; BA: bone age; US: ultrasonography.

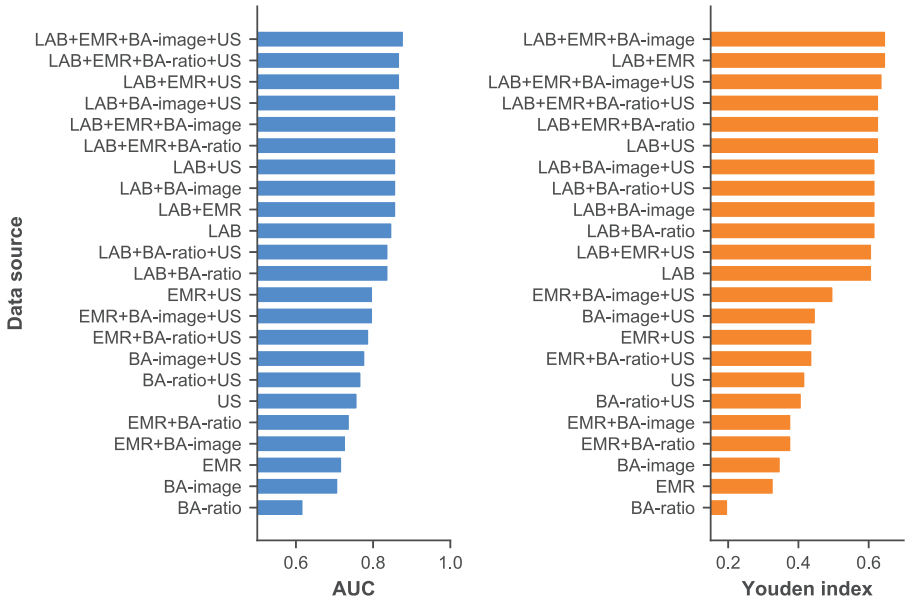

Figure 3. AUC and Youden index ranking among 23 models. The left panel represents the AUC of the models based on different data source combinations, and the right panel represents the Youden index. LAB: laboratory; EMR: electronic medical records; BA: bone age; US: ultrasonography.

GnRH stimulation test and identify girls with CPP in a timely manner.

Generally, multisource data fusion in the medical field aims at the fusion of single modal data such as the combination of laboratory results and demographic characteristics. Nowadays, medical image examination usually becomes a routine auxiliary procedure for diagnosis. Features extracted from images by deep learning technologies may contain more information than examination reports written by radiologists. However, one challenge for multisource and heterogeneous data fusion in clinics is how to effectively utilize different data types from different data sources. Here, we tried to fuse multisource and heterogeneous data on the feature level. The data we used were produced during a whole diagnostic process and were supposed to contain more information than any single-source data. The success was demonstrated by comparisons between single-source and multisource models. Studies in other fields have also indicated that multimodal information can provide complementary information to improve model performance.[14,22]

From the importance ranking of the top 20 features, basal LH presented high diagnostic value, followed by uterine volume. Sathasivam et al[23] compared the ovarian and uterine volumes from pelvic ultrasonography with the basal and stimulated LH to assess girls with suspected precocious puberty. The study found that basal LH and stimulated LH significantly correlated with ovarian and uterine volumes. Our study indicated that pelvic ultrasonography alone cannot differentiate girls with CPP from non-CPP well (sensitivity of 66.23%). But once combining with the LAB data, the model showed a greatly improved Youden index from 0.42 to 0.63.

Compared to our previous study which obtained an AUC of 0.90, the models in this study only yielded an AUC as high as 0.88. However, the AUC in the previous study was an average value of 10-fold crossvalidation using data of 436 patients, while the performance in this study was tested on an independent test set of 180 patients. Here, we also trained the best XGBoost model on the whole set of 900 patients through 10-fold crossvalidation, and got an average AUC of 0.93. This indicates that multisource data fusion does help to improve CPP diagnosis.

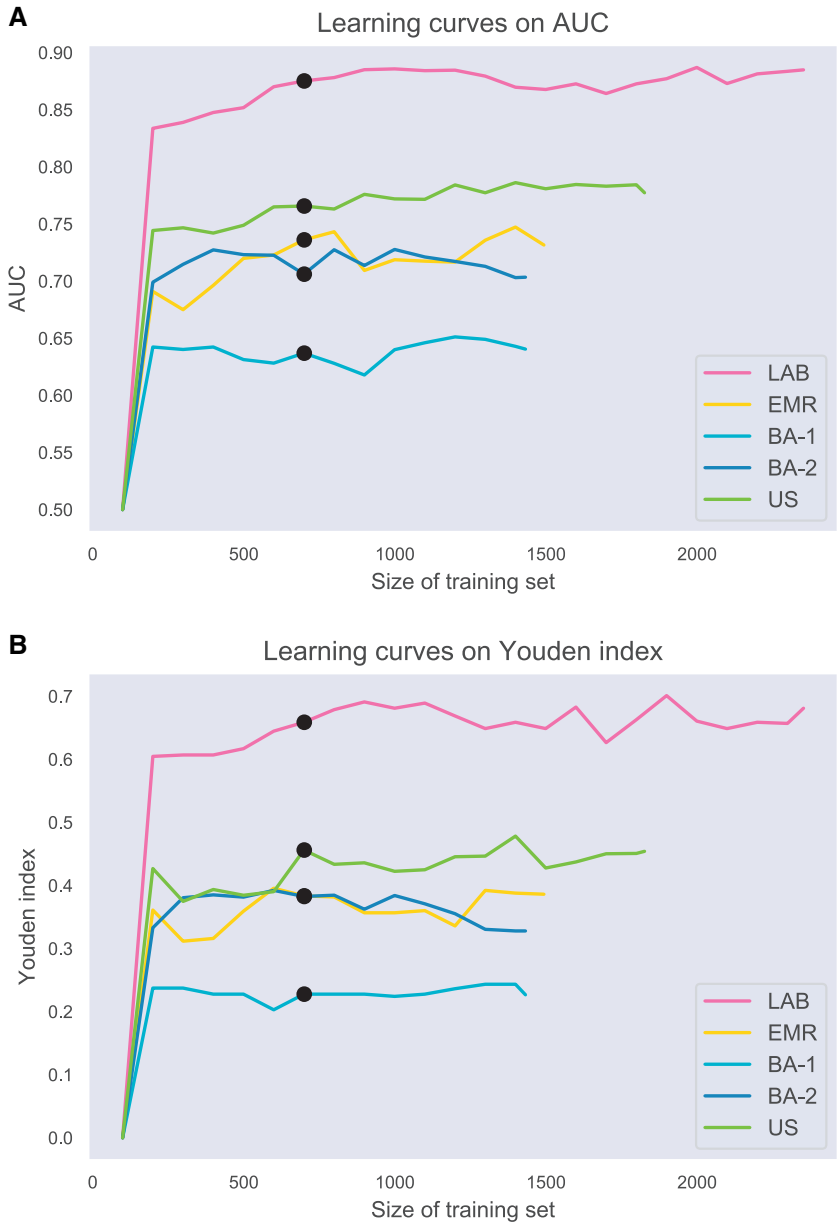

Figure 4. Learning curves for models based on different single-source data. The black dot represents the performance at the maximum size ($n = 720$) of the training set with comprehensive data. The horizontal axis represents the sample size, and the vertical axis represents the model performance that varied with the sample size. LAB: laboratory; EMR: electronic medical records; BA: bone age; US: ultrasonography.

The application of image analysis based on deep learning techniques and machine learning algorithms is innovative in CPP diagnosis and eliminates the subjectivity of BA value assessment. BA X-ray images have recently been used in BA value assessment with deep learning models.[24,25] In this study, we extracted features from BA images using a deep learning model. These features represented skeletal maturity. Of the top 20 most important features, 60% are the features with the prefix "Img-," that is, those image features that are unexplainable and not previously discovered. This reveals that artificial intelligence may explore important features that human beings may leave out.

Among the models developed in this study, we selected three simplified models tailored for patients in different hospitals (see Supplementary Figure S1). Thus, endocrinologists can focus on appropriate variables with high diagnostic value, especially for resource-limited settings, to get a fast and accurate decision. For example, in a tertiary children's hospital or general hospital with equipments to perform sexual hormone tests, a model constructed by data from all four sources can achieve good performance (Supplementary Figure S1(B)). Before performing the examinations, focusing on information related to secondary sexual development and basal serum hormone test can also contribute to making a preliminary decision (Supplementary Figure S1(A)). In nontertiary hospitals with the sexual hormone test unavailable, using a model built by data from the EMR, BA image, and US sources will provide a preliminary result to screen CPP (Supplementary Figure S1(C)). Instead of considering the overall factors and the GnRH stimulation test, focusing on the optimal feature set with fewer factors without a stimulation test is efficient and convenient for physicians and patients.

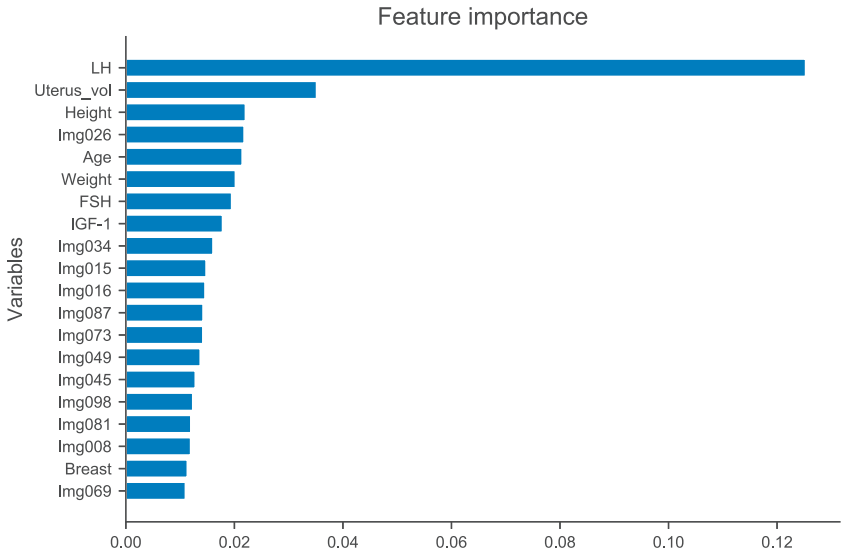

Figure 5. Feature importance of each variable (top 20). LH: luteinizing hormone; FSH: follicle-stimulation hormone; IGF-1: insulin-like growth factor-1.

To use as much training data as possible to improve the model performance, we considered the relationship between sample size and model performance. This analysis suggested that with more training data, the performance of each data source improved. The collection of more patients with comprehensive data from multiple data sources will result in a more precise diagnosis for girls with suspected CPP. Although the three models did not reach 100% sensitivity or specificity as the stimulation test did, we envision it as a guide or supplementary tool in CPP screening.

Our study has several limitations. First, as we considered utilizing multisource data to build models, the eligible sample size decreased, which is inevitable in retrospective studies. From the learning curves based on data size, a larger sample size will make single-source data work better, and the same is true with the multisource data. Second, there may exist a time interval between the earlier data collection and the data collection from the GnRHa stimulation test. Sometimes, the interval between a BA examination and the stimulation test could be a few months. A patient with BA value assessed a few months ago may have had an accelerated bone growth by the time she has a positive response in the stimulation test. These changes may have had more or less influence on model performance. It would be better to have a patient's comprehensive set of data at the same time, or at least with the shorter time interval possible. Third, all the results were achieved on the independent data set comprised of 180 patients with data from all the four sources. This provides evidence for model generalizability; however, further the evaluation between facilities at separate locations or a prospective study would provide stronger support of model generalization.

CONCLUSION

In conclusion, we developed machine learning models using easily accessed clinical data without the inconvenient stimulation test to identify girls who are at high risk for CPP among patients with precocious puberty. Three simplified diagnostic methods are tailored to the needs of patients in different clinical settings. The diagnostic value of basal laboratory parameters is of high significance in CPP diagnosis. The models with multisource and heterogeneous data performed better than the models with single-source data.

AUTHORS' CONTRIBUTIONS

HL proposed the project and provided supervision. LP is responsible for data collection, analysis, manuscript drafting and revision. GL implemented part of data analysis and performed manuscript editing. XM provided clinical guidance and revised the manuscript. All authors reviewed the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

FUNDING

This work was supported by National Key Research and Development Project of China [grant number 2018YFC1315402] and Guangzhou Institute of Pediatrics, Guangzhou Women and Children's Medical Center [grant number YIP-2019-064, IP-2019-017].

CONFLICT OF INTEREST STATEMENT

The authors have declared that these is no conflict of interest.

DATA AVAILABILITY

The data underlying this article will be shared on reasonable request to the corresponding author.

REFERENCES

1. Sultan C, Gaspari L, Kalfa N, *et al. Clinical Expression of Precocious Puberty in Girls. Pediatric and Adolescent Gynecology*. Berlin: Karger Publishers; 2012: 84–100.
2. Fisher MM, Eugster EA. What is in our environment that effects puberty? *Reprod Toxicol* 2014; 44: 7–14.
3. Kim YJ, Kwon A, Jung MK, *et al.* Incidence and prevalence of central precocious puberty in Korea: an epidemiologic study based on a national database. *J Pediatr* 2019; 208: 221–8.

4. Abreu AP, Kaiser UB. Pubertal development and regulation. *Lancet Diabetes Endocrinol* 2016; 4 (3): 254–64.
5. Houk CP, Kunselman AR, Lee PA. The diagnostic value of a brief GnRH analogue stimulation test in girls with central precocious puberty: a single 30-minute post-stimulation LH sample is adequate. *J Pediatr Endocrinol Metab* 2008; 21 (12): 1113–8.
6. Kandemir N, Demirbilek H, Özön ZA, *et al.* GnRH stimulation test in precocious puberty: single sample is adequate for diagnosis and dose adjustment. *J Clin Res Pediatr Endocrinol* 2011; 3 (1): 12–7.
7. Yazdani P, Lin Y, Raman V, *et al.* A single sample GnRHa stimulation test in the diagnosis of precocious puberty. *Int J Pediatr Endocrinol* 2012; 2012 (1): 23.
8. Ding Y, Li J, Yu Y, *et al.* Evaluation of basal sex hormone levels for activation of the hypothalamic-pituitary-gonadal axis. *J Pediatr Endocrinol Metab* 2018; 31 (3): 323–29.
9. Lee SH, Joo EY, Lee J-E, *et al.* The diagnostic value of pelvic ultrasound in girls with central precocious puberty. *Chonnam Med J* 2016; 52 (1): 70–74.
10. De Vries L, Phillip M. Role of pelvic ultrasound in girls with precocious puberty. *Horm Res Paediatr* 2011; 75 (2): 148–52.
11. Eksioglu AS, Yilmaz S, Cetinkaya S, *et al.* Value of pelvic sonography in the diagnosis of various forms of precocious puberty in girls. *J Clin Ultrasound* 2013; 41 (2): 84–93.
12. Lee HK, Choi SH, Fan D, *et al.* Evaluation of characteristics of the craniofacial complex and dental maturity in girls with central precocious puberty. *Angle Orthod* 2018; 88 (5): 582–9.
13. Pan L, Liu G, Mao X, *et al.* Development of prediction models using machine learning algorithms for girls with suspected central precocious puberty: retrospective study. *JMIR Med Inform* 2019; 7 (1): e11728.
14. Multimodal deep learning for cervical dysplasia diagnosis. In: International conference on medical image computing and computer-assisted intervention. 2016: 115–23; Cham: Springer.
15. Yap J, Yolland W, Tschandl P. Multimodal skin lesion classification using deep learning. *Exp Dermatol* 2018; 27 (11): 1261–67.
16. Xie Y, Zhang J, Xia Y, *et al.* Fusing texture, shape and deep model-learned information at decision level for automated classification of lung nodules on chest CT. *Information Fusion* 2018; 42: 102–10.
17. ICML. In: JMLR Proceedings 27, JMLR.org 2012; July 2, 2011; Bellevue, Washington.
18. Subspecialty Group of Endocrinologic, Hereditary and Metabolic Diseases, the Society of Pediatrics, Chinese Medical Association; Editorial Board, Chinese Journal of Pediatrics. Consensus statement for the diagnosis and treatment of central precocious puberty (2015). *Zhonghua Er Ke Za Zhi* 2015; 53 (6): 412–8.
19. Carel J-C, Eugster EA, Rogol A, on behalf of the members of the ESPE-LWPES GnRH Analogs Consensus Conference Group, *et al.* Consensus statement on the use of gonadotropin-releasing hormone analogs in children. *Pediatrics* 2009; 123 (4): e752–e62.
20. Liu X, Wang G, Xu Y, *et al.* Bone age assessment by deep convolutional neural networks combined with clinical TW3-RUS. In: 2019 IEEE international conference on bioinformatics and biomedicine (BIBM); 2019.
21. KDD '16. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; San Francisco, CA; August 2016; New York: ACM.
22. Rajalingam B, Priya R. Multimodal medical image fusion based on deep learning neural network for clinical treatment analysis. *Int J Chemtech Res* 2018; 11 (06): 160–76.
23. Sathasivam A, Rosenberg HK, Shapiro S, *et al.* Pelvic ultrasonography in the evaluation of central precocious puberty: comparison with leuprolide stimulation test. *J Pediatr* 2011; 159 (3): 490–95.
24. Spampinato C, Palazzo S, Giordano D, *et al.* Deep learning for automated skeletal bone age assessment in X-ray images. *Med Image Anal* 2017; 36: 41–51.
25. Iglovikov VI, Rakhlin A, Kalinin AA, *et al.* Paediatric bone age assessment using deep convolutional neural networks. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer; 2018: 300–08.