A Classification Tool for Differentiation of Kawasaki Disease from Other Febrile Illnesses

Shiying Hao, PhD[1], Bo Jin, MS[1], Zhou Tan, PhD[1], Zhen Li, BS[1], Jun Ji, PhD[1], Guang Hu, PhD[1], Yue Wang, BS[1], Xiaohong Deng, PhD[1], John T. Kanegaye, MD[2,3], Adriana H. Tremoulet, MD, MAS[2,3], Jane C. Burns, MD[2,3], Harvey J. Cohen, MD, PhD[4], and Xuefeng B. Ling, PhD[1], on behalf of the Pediatric Emergency Medicine Kawasaki Disease Research Group*

Objective To develop and validate a novel decision tree-based clinical algorithm to differentiate Kawasaki disease (KD) from other pediatric febrile illnesses that share common clinical characteristics.

Study design Using clinical and laboratory data from 801 subjects with acute KD (533 for development, and 268 for validation) and 479 febrile control subjects (318 for development, and 161 for validation), we developed a step-wise KD diagnostic algorithm combining our previously developed linear discriminant analysis (LDA)–based model with a newly developed tree-based algorithm.

Results The primary model (LDA) stratified the 1280 subjects into febrile controls (n = 276), indeterminate (n = 247), and KD (n = 757) subgroups. The subsequent model (decision trees) further classified the indeterminate group into febrile controls (n = 103) and KD (n = 58) subgroups, leaving only 29 of 801 KD (3.6%) and 57 of 479 febrile control (11.9%) subjects indeterminate. The 2-step algorithm had a sensitivity of 96.0% and a specificity of 78.5%, and correctly classified all subjects with KD who later developed coronary artery aneurysms.

Conclusion The addition of a decision tree step increased sensitivity and specificity in the classification of subject with KD and febrile controls over our previously described LDA model. A multicenter trial is needed to prospectively determine its utility as a point of care diagnostic test for KD. *(J Pediatr 2016; ■: ■-■).*

More effective methods for the early diagnosis of acute Kawasaki disease (KD) are required to permit timely administration of intravenous immunoglobulin and prevention of adverse outcomes. The classic KD diagnostic criteria adopted by the American Heart Association (AHA) include fever plus ≥4 of 5 principal clinical signs (**Figure 1**).[1] These guidelines, although widely adopted by clinicians, occasionally fail to differentiate KD from other pediatric rash/fever illnesses.[2] Moreover, despite supplementary laboratory criteria to aid in the diagnosis of patients with KD who manifest only 2 or 3 clinical signs, these incomplete cases may still be missed by clinicians.[1] Missing the diagnosis can lead to delayed treatment, thus increasing the risk of developing coronary artery lesions.[3-5]

We previously applied statistical learning using clinical and laboratory test variables, and developed a linear discriminant analysis (LDA)-based scoring system to differentiate KD from febrile controls[6] with a sensitivity of 92%-94% and a specificity of 88%-89%. However, 20%-30% of subjects in either the KD or febrile controls groups remained unclassified, and the algorithm performance on subjects with KD with incomplete clinical criteria was not investigated.[6,7]

In this study, we tested the hypothesis that applying separate tree-based algorithms after the LDA algorithm would improve the classification accuracy in differentiating subjects with KD from febrile control subjects. This novel integrated algorithm was validated with an independent subject cohort.

Methods

Subjects with KD and febrile controls meeting inclusion criteria were identified from the database maintained at the University of California San Diego KD

AHA	American Heart Association
CRP	C-reactive protein
ED	Emergency department
KD	Kawasaki disease
LAD	Left anterior descending
LDA	Linear discriminant analysis
NPV	Negative predictive value
PPV	Positive predictive value
RCA	Right coronary artery

From the [1]Department of Surgery, Stanford University, Stanford; [2]Department of Pediatrics, University of California San Diego, La Jolla; [3]Rady Children's Hospital San Diego, San Diego; [4]Department of Pediatrics, Stanford University, Stanford, CA

*List of additional members of the Pediatric Emergency Medicine Kawasaki Disease Research Group is available at www.jpeds.com (**Appendix 1**).

Supported by the American Heart Association (to H.C. and X.L.), Stanford University Spark Program (H.C. and X.L.), the David Gordon Louis Daniel Foundation (to J.B.), the Mario Batali Foundation (J.B.), the National Institutes of Health, National Heart, Lung, Blood Institute (HL69413 [to J.B.]), the Hartwell Foundation (to A.T.), and the Harold Amos Medical Faculty Development Program/Robert Wood Johnson Foundation (to A.T.). The authors declare no conflicts of interest.

http://dx.doi.org/10.1016/j.jpeds.2016.05.060

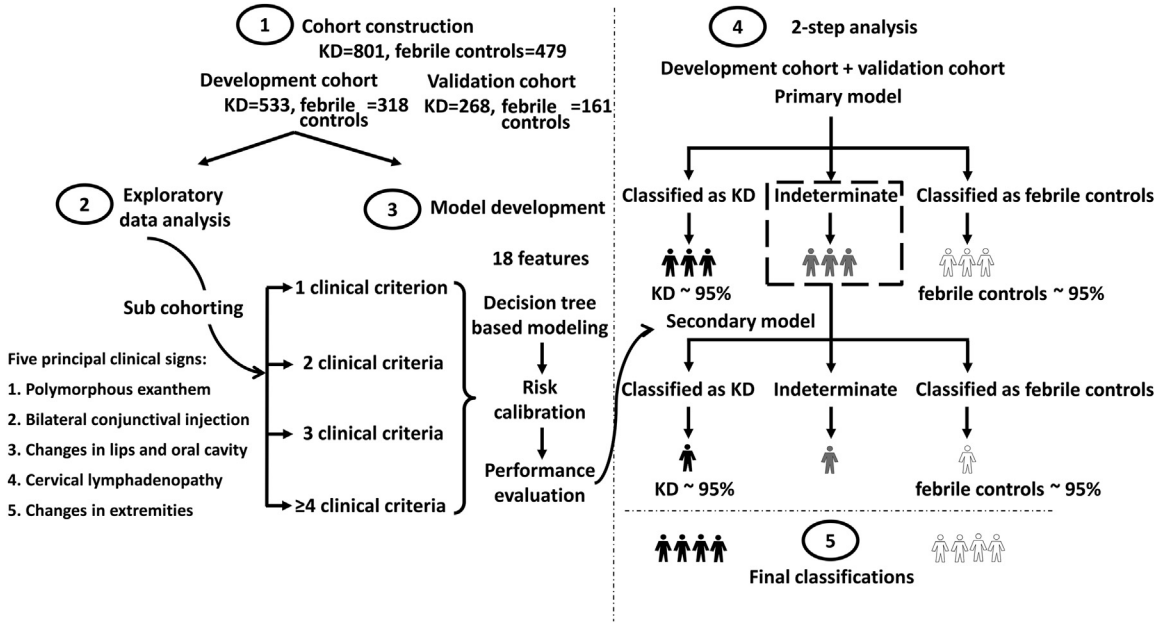


Figure 1. Workflow to create a 2-step statistical algorithm for distinguishing subjects with KD and febrile control subjects. LDA- and decision-tree–based models developed based on clinical and laboratory test variables were applied in sequence to construct a 2-step algorithm, partitioning the subjects into 3 diagnostic classifications (febrile controls, KD, and indeterminate). PPV and NPV of 95% were achieved at each step.

Research Center. Complete demographic and clinical data were collected prospectively on all subjects with KD and febrile controls. A total of 1280 subjects (801 with KD and 479 febrile controls) were included in this study (**Figure 2**; available at www.jpeds.com). Subjects with KD in this study were (a) patients with fever (≥38.0°C rectally or orally) for no more than 10 days plus ≥4 of the 5 principal clinical criteria, (b) patients meeting fewer criteria but with coronary artery abnormalities (*Z*-score ≥ 2.5 for left anterior descending [LAD] and/or right coronary arteries [RCA]) documented by echocardiography, and (c) patients meeting <4 criteria but meeting the AHA criteria for incomplete KD by laboratory criteria.[1] A concomitant viral infection by reverse transcriptase polymerase chain reaction did not disqualify the patient as a KD subject. Every subject was evaluated clinically by 1 of 2 expert KD clinicians and the final assignment of a KD diagnosis was based on the opinion of these 2 experts. Febrile control subjects were recruited from the emergency department (ED) at Rady Children's Hospital San Diego. All febrile control subjects had unexplained fever, ≥1 of the 5 principal clinical criteria for KD, and had laboratory tests performed including those commonly ordered for evaluation of KD, which included a complete blood count with manual differential, erythrocyte sedimentation rate, and levels of C-reactive protein (CRP), alanine aminotransferase, and gamma glutamyl transferase. All patients referred to the ED for evaluation of possible KD (approximately 50% of the febrile control cohort) were offered enrollment as febrile control subjects in our study. We enrolled the remaining febrile controls from children in the ED presenting with fever and ≥1 of the clinical signs of KD, and excluded patients who had an obvious respiratory or gastrointestinal infection because KD would be unlikely to present in this manner. The final diagnoses of the febrile controls were determined by chart review by 2 expert clinicians from prospectively collected clinical and laboratory data and from review of microbiologic and serologic results and subsequent clinical encounters. Only 3.8% of the febrile controls (18 of 479) underwent echocardiography to evaluate for possible KD.

Signed consent or assent forms were obtained from the parents of all subjects and from all subjects >6 years of age. The study was approved by the institutional review boards of the University of California San Diego and Stanford University.

For each subject, we collected the 18 clinical and laboratory test variables retained in the final model of the LDA-based algorithm.[6] Clinical data included 6 clinical signs associated with KD: illness days (temperature ≥38.0°C); cervical lymph node of ≥1.5 cm; rash; conjunctival injection; extremity changes including red, swollen, or peeling hands or feet; and oropharyngeal changes including red pharynx, red, fissured lips, or strawberry tongue. Laboratory test data (obtained prior to administration of intravenous immunoglobulin for subjects with KD) included total white blood cell

count, percentages of monocytes, lymphocytes, eosinophils, neutrophils, and immature neutrophils (bands), platelet count, hemoglobin concentration normalized for age, CRP, gamma glutamyl transferase, alanine aminotransferase, and erythrocyte sedimentation rate. For test results that exceeded the upper or lower limit of the test, we used the numeric value for the limit. Subgroups of subjects with KD were defined as having either normal coronary arteries (RCA and LAD Z-score always <2.5), transiently dilated coronary arteries (RCA and/or LAD Z-score ≥2.5 and resolving within 8 weeks of KD onset), or aneurysms (Z-score ≥5.0 or dilated segment 1.5 times the internal diameter of the adjacent segment). We performed a multivariate analysis on the clinical and laboratory test variables for KD and febrile controls discrimination in our total dataset. Panels combining 18 clinical and laboratory test variables were evaluated and the resulting ORs, P values, and variable effects on the final model were calculated.

Of the 1280 subjects (801 KD and 479 febrile controls), 489 (261 KD and 228 febrile controls) were from the development cohort in our previous study[6] and remained in this study's development cohort. The remaining 791 subjects were assigned into 2 cohorts while maintaining the same ratio of KD and febrile controls subjects across cohorts. Of the entire cohort of 1280 subjects, 228 of 801 subjects with KD and 287 of 479 febrile controls subjects had ≥1 missing value for the laboratory variables. Missing values were imputed among KD and febrile controls subjects, respectively, using a method of weighted K-nearest neighbors (**Appendix 2**; available at www.jpeds.com).[8] There were 533 subjects with KD and 318 febrile controls for model development, and 268 subjects with KD and 161 febrile controls for model validation. The study design is outlined in **Figure 1**. A 2-step algorithm was developed using the 6 clinical and 12 laboratory test variables to stratify the subjects into 3 subgroups: febrile controls, indeterminate, and KD. The 95% positive predictive value (PPV) and negative predictive value (NPV) for KD and febrile controls classification were targeted at each step.

Primary Model

The previously developed KD algorithm was developed using an LDA method, with days of fever, 5 principal clinical criteria, and 12 laboratory test variables as input variables. The output of the algorithm was a unique score describing the probability of KD diagnosis for each subject.[6] Two cutoffs were set to stratify these subjects into 3 classification subgroups: febrile controls, indeterminate, and KD,[6] allowing 95% accuracy in both KD and febrile controls subgroups.

After applying the LDA model, 9.6% of subjects with KD (51 of 533) and 33.0% of febrile control subjects (105 of 318) in the development cohort remained indeterminate. The proportions of subjects with indeterminate scores, however, differed among the 4 subcohorts based on the number of principal clinical criteria manifested by each subject. The LDA model performed less well for subjects with fewer clinical criteria, yielding indeterminate scores for 28.4% of subjects with KD (29 of 102) and 43.9% of febrile controls (72 of 164) who manifested only 2 or 3 clinical criteria. Therefore, an additional model was developed to improve the adjudication of indeterminate subjects based on the number of clinical criteria present.

Secondary Model

To improve the classification of subjects in the indeterminate group from the first analysis, we used 2-step data mining methods to combine the advantages of multiple models to achieve better predictive accuracy than is possible with any individual model.[9] Random forest models constructed by a set of decision trees were developed.[10,11] Subjects were divided into 4 subcohorts based on the number of KD criteria that they manifested (**Figure 1**). Separate models were then developed for each subcohort. Specifically, subjects in the development cohort were further randomly partitioned into 2 subcohorts (subcohort I and subcohort II). A 'forest' of 300 binary 'trees' was constructed using randomly selected samples and variables (clinical and laboratory test variables) of subcohort I. At each node, 'trees' were split by choosing a split variable value producing the maximum node separation. 'Trees' were constructed until each of the terminal nodes reached a sample size of 1. Final decisions were reached by averaging the decisions of each tree. The derived algorithm was then calibrated with subcohort II by setting 2 thresholds that stratified all the subjects into 3 classification subgroups (febrile controls, indeterminate, and subjects with KD), allowing 95% PPV and NPV. The performance of the algorithm was tested on the validation cohort. The modeling details appear in **Appendix 3** (available at www.jpeds.com).

Performance Analyses

Performance of the 2-step model was demonstrated by sensitivity, specificity, PPV, and NPV. Classification of incomplete subjects with KD and subjects developing coronary artery abnormalities was analyzed. Indeterminate subjects were analyzed to explore the model limitations. Performance of models derived with reduced numbers of input variables (missing data) was tested to explore its robustness in KD/febrile controls classification.

Results

The demographic and clinical details of development and validation cohorts are presented in **Table I**. Asian patients were overrepresented among subjects with KD and underrepresented among febrile controls, compared with the San Diego population at large (12%). Febrile controls had a clinically determined or culture-proven etiology for their febrile illnesses (**Table II**; available at www.jpeds.com). Viral diagnosis was established by viral culture, direct fluorescent antibody testing, or polymerase chain reaction assays. Viral syndrome was defined as a febrile illness that resolved without specific treatment and for which no specific pathogens could be identified.

Table I. Demographic characteristics of study cohorts

Characteristics	Development cohort			Validation cohort		
	KD (n = 533)	Febrile controls (n = 318)	*P*	KD (n = 268)	Febrile controls (n = 161)	*P*
Age, mo, median (IQR)	29.8 (15.8, 52.0)	30.7 (15.4, 61.8)	.22*	30.4 (16.8, 52.6)	45.0 (18.9, 79.1)	<.001*
Males, n (%)	337 (63.2)	191 (60.1)	.38†	157 (58.6)	98 (61)	.69†
Race/ethnicity, n (%)			<.001†			.03†
African American	22 (4.1)	8 (2.5)		11 (4.1)	3 (2)	
Native American	2 (0.4)	0 (0)		1 (0.4)	0 (0)	
Asian	91 (17.1)	26 (8.2)		45 (16.8)	11 (7)	
Caucasian	120 (22.5)	83 (26.1)		72 (26.9)	45 (28)	
Hispanic	175 (32.8)	124 (39.0)		84 (31.3)	59 (37)	
Mixed	109 (20.5)	60 (18.9)		45 (16.8)	32 (20)	
Other/unknown	14 (2.6)	17 (5.3)		10 (3.7)	11 (7)	

*Rank sum test.
†Fisher exact test.

Multivariate Analysis and 2-Step Analyses of KD and Febrile Controls

We compared subject with KD and febrile controls using the Fisher exact tests for categorical variables, and ORs and likelihood ratio tests for continuous variables (**Tables III** and **IV**; available at www.jpeds.com). Each subcohort had different statistically significant clinical variables in the univariate analysis and independent predictors in the multivariate analysis (**Tables III** and **IV**), supporting the need to develop models for each subcohort separately. The impacts of each variable to the classification decision in secondary models were measured by the percent increase of model mean square error owing to the permutation of the variable values (**Table V**; available at www.jpeds.com).

By applying the previously derived primary LDA-based model and the score cutoffs that achieved 95% PPV and NPV,[6] 90.0% of subjects with KD (721 of 801) and 57.4% of febrile controls (275 of 479) were correctly classified; 0.1% of subjects with KD (1 of 801) and 7.5% of febrile controls (36 of 479) were erroneously classified, and 9.9% of subjects with KD (79 of 801) and 35.1% of febrile controls (168 of 479) were left indeterminate.

The secondary random forest models, applied to 4 subcohorts of remaining indeterminate subjects, correctly classified 60.8% of subjects with KD (48 of 79) and 60.1% of febrile controls (101 of 168). The secondary models erroneously classified 2.5% of subjects with KD (2 of 79) and 6.0% of febrile controls (10 of 168), and 36.7% subjects with KD (29 of 79) and 33.9% of febrile controls (57 of 168) remained indeterminate.

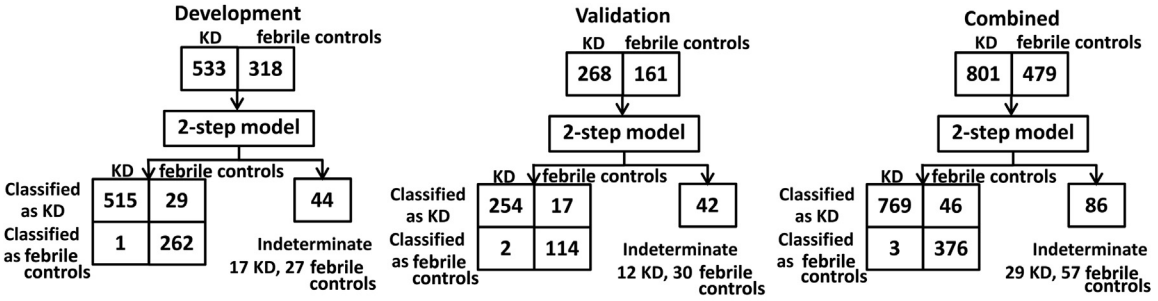
The 2-step algorithm correctly classified 96.0% of subjects with KD (769 of 801) and 78.5% of febrile controls (376 of 479; **Figure 3**) with targeted ≥95% PPV and NPV. Only 3.6% of subjects with KD (29 of 801) and 11.9% of febrile controls (57 of 479) remained indeterminate, whereas 9.9% of subjects with KD and 35.1% of febrile controls subjects were left indeterminate by the original LDA model.

We compared the ability of the 2-step algorithm in terms of sensitivity, specificity, PPV, and NPV to the use of the AHA guidelines for KD diagnosis in the absence of echocardiography (**Figure 4**; available at www.jpeds.com). Results showed that the algorithm had a sensitivity of 96.0% vs AHA guidelines of only 72.2%. The AHA guidelines had a higher specificity of 93.5% vs our specificity of 78.5%. However, when it came to PPV and NPV, the AHA guidelines and the 2-step algorithm had the same PPV around 95%, and our NPV was 99.2% whereas AHA guidelines had a NPV of 66.8%. Thus, use of the algorithm was better at identifying more patients with KD and having a better NPV for patients without KD.

Algorithm Performance in Subcohorts Stratified by Age, Illness Day, and CRP

The diagnosis of KD in young infants can be particularly challenging. This algorithm performed well (**Table VI**; available at www.jpeds.com) among subjects ≤6 months of age (n = 92; 69 subjects with KD and 23 febrile controls). The PPV and NPV for these infants were both 100%. The sensitivity was 97.1% and specificity was 87.0%. Only 2 subjects with KD (3%) and 3 febrile controls (13%) were indeterminate. Among subjects >6 months of age (n = 1188; 732 subjects with KD and 456 febrile controls), the PPV, NPV, sensitivity, and specificity were slightly lower (93.9%, 99.2%, 95.9%, and 78.1%, respectively). The indeterminate frequency of these older subjects with KD and febrile controls were 3.7% (27 of 732) and 11.8% (54 of 456). The distribution of correctly classified, erroneously classified, and indeterminate subjects did not differ significantly among the 2 age groups ($P = .12$ by χ^2 test). Importantly, the algorithm performed well in different age groups including the most vulnerable age group, namely, patients <6 months of age.

To determine the effect of duration of illness on algorithm performance, we divided the subjects into 4 subcohorts based on illness day (≤3 days [n = 251]; 4-5 days [n = 435]; 6-7 days [n = 390]; 8-10 days [n = 204]) and analyzed the algorithm's performance for each subcohort (**Table VI**). PPVs, NPVs, and sensitivities remained similar (<7% variation) among these subcohorts. The specificity levels decreased monotonically with illness duration from 85.7% the group of ≤3 days of illness to 61.2% the subcohort of 8-10 days of illness. The distribution of correctly classified, erroneously classified, and indeterminate subjects did not differ among subcohorts of illness days ($P = .27$ by χ^2 test).



Cohort	Sensitivity[a] and 95% CI (%)	Specificity[a] and 95% CI (%)	PPV and 95% CI (%)	NPV and 95% CI (%)	Indeterminate KD after step 1 (%)	Indeterminate febrile controls after step 1 (%)	Indeterminate KD after step 2 (%)	Indeterminate febrile controls after step 2 (%)
Development	96.6 (94.6, 97.9)	82.4 (77.7, 86.3)	94.7 (92.3, 96.3)	99.6 (97.6, 100)	9.6	33.0	3.2	8.5
Validation	94.8 (91.2, 97.0)	70.8 (63.0, 77.6)	93.7 (90.0, 96.2)	98.3 (93.3, 99.7)	10.4	39.1	4.5	18.6
Combined	96.0 (94.3, 97.2)	78.5 (74.5, 82.0)	94.4 (92.5, 95.8)	99.2 (97.5, 99.8)	9.9	35.1	3.6	11.9

[a] With targeted ≥ 95% PPV and NPV

Figure 3. Diagnostic performance of the 2-step algorithm applied to the development and validation cohorts. *Top*, Classification of subjects. *Bottom*, Sensitivity, specificity, PPV, NPV, and proportions of subjects with indeterminate scores.

In our study, there were 242 of 479 febrile controls (50.5%) who had CRP values of ≥3.0 mg/dL. Of these febrile controls, the algorithm correctly classified 72.7% of the subjects (176 of 242), erroneously classified 12.8% of the subjects (31 of 242), and left 14.5% of the subjects (35 of 242) indeterminate. Of these 242 subjects, 61 fulfilled the criteria for incomplete KD based on AHA guidelines. This algorithm correctly identified 42.6% as febrile controls (26 of 61) and left 41.0% as indeterminate subjects (25 of 61) requiring further evaluation. For the 237 febrile control subjects who had CRP of <3.0 mg/dL, the algorithm classified 84.4% (n = 200) correctly as febrile controls, 6.3% (n = 15) erroneously as KD, and 9.3% (n = 22) as indeterminate. Such results demonstrate the utility of the algorithm as a classification tool for frontline clinicians to evaluate suspected KD when echocardiography is not readily available.

Algorithm Performance for Subjects with Incomplete KD

Of 801 subjects with KD, 646 had complete KD, 155 met AHA criteria for incomplete KD with 62 showing coronary changes (57 had transiently dilated coronary arteries and 5 had aneurysms) on the initial echocardiogram. For the 93 incomplete subjects with KD with normal echocardiograms, the algorithm classified 80.6% (n = 75) correctly as KD, 1.1% (n = 1) erroneously as febrile controls, and 18.3% (n = 7) as indeterminate. Compared with the original LDA model (26.9% indeterminate), the 2-step model improved the correct adjudication of incomplete KD by almost one-third.

Classification of KD with Coronary Artery Abnormalities

Because the prompt diagnosis of the subset of subjects with KD who developed coronary artery aneurysms is of paramount importance, the model's performance was separately evaluated for subjects in regard to coronary artery status (Figure 5). Of the 32 subjects with KD who were classified erroneously or indeterminate, 26 had normal and 6 had transiently dilated

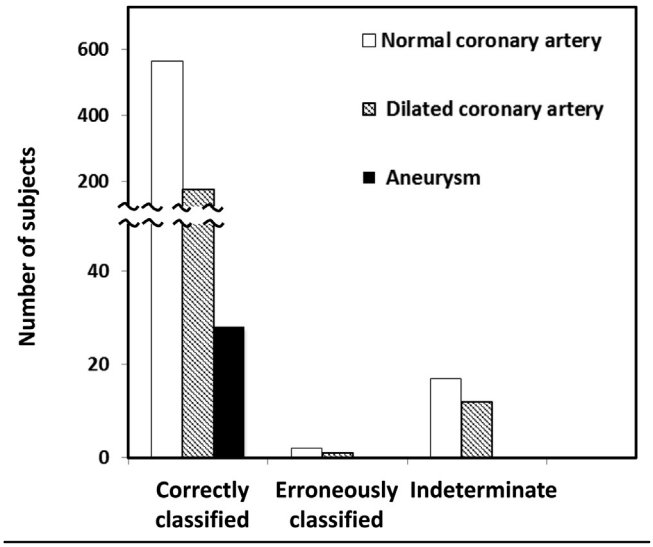


Figure 5. Performance of the algorithm according to coronary artery status of subjects.

coronary arteries (with worst *Z*-scores ranging from 2.66 to 4.35 for the LAD and/or RCA). Thus, the algorithm correctly classified all 28 subjects who developed aneurysms based on baseline clinical criteria and laboratory test results before the echocardiography was performed. The distribution of these 28 subjects over the 3 subcohorts was shown in **Figure 6** (available at www.jpeds.com). Five of these 28 subjects manifested 2 or 3 criteria and were diagnosed by echocardiography. In addition, 57 subjects who manifested 2 or 3 criteria had dilation of the coronary arteries (*Z*-score > 2.5) and were diagnosed on this basis. The decision support tool correctly identified 81% of subjects with incomplete KD (50 of 62) diagnosed by echocardiographic criteria. Thus, the decision support tool could be used on the initial examination to improve the diagnosis of KD.

Erroneously Classified and Indeterminate Subjects

Clinical and laboratory test variables were analyzed to profile the subjects with erroneous or indeterminate classification by the model. Of the 32 erroneously classified or indeterminate subjects with KD, 30 manifested ≤3 KD principal criteria. Thus, the majority exhibited incomplete clinical characteristics at the time that the algorithm was applied. In contrast, 70 of 103 erroneously classified or indeterminate febrile control subjects manifested ≥3 KD principal criteria. Distributions of the 12 laboratory test variables were compared among the correctly classified, erroneously classified, and indeterminate KD and febrile control subgroups (**Figure 7**; available at www.jpeds.com). The subjects with KD erroneously classified as febrile controls had laboratory test values comparable with those of correctly classified febrile controls. Conversely, the laboratory test values of indeterminate subjects with KD and febrile controls were intermediate to those of the correctly classified KD and febrile control subjects. Adenovirus is well-known to mimic many of the clinical and laboratory features of KD.[12] For the 28 febrile control subjects with adenovirus infection documented either by culture, direct fluorescent antibody testing, or polymerase chain reaction in the validation cohort, the algorithm classified 57% (n = 16) correctly as febrile controls, 18% (n = 5) erroneously as having KD, and 25% (n = 7) as indeterminate. Such clinical and laboratory test result patterns likely explain the misclassification by the algorithm.

Impact of Variable Reduction in Algorithm Performance

Because patients typically have incomplete data early in their evaluations, we studied the effect of eliminating variables, beginning with the least weighted (**Figure 8**; available at www.jpeds.com). The frequency of certain classifications decreased with the reduction in variable number from 18 to 3. A 9-variable algorithm including 6 clinical variables (5 KD principal criteria plus illness days) and 3 laboratory variables (hemoglobin concentration normalized for age, eosinophil percentage, and white blood cell count) had an 80% classification certainty rate for subjects with KD and febrile controls and 42% for incomplete subjects with KD. In our study cohort, there were 228 of 801 subjects with KD and 287 of 479 febrile controls with missing laboratory values. The impact of these subjects on algorithm performance was also explored (**Appendix 2** and **Table VII**; **Table VII** available at www.jpeds.com). Missing laboratory data did not affect the algorithm performance and both the negative and PPV was preserved (**Table VII**).

Discussion

The sequential use of a primary LDA-derived algorithm to perform initial classification and secondary decision-tree-based algorithms applied in parallel to subcohorts of indeterminate cases resulted in improved classification certainty in differentiating KD from clinically similar febrile illnesses. The diagnosis of patients with incomplete KD criteria is also challenging. The algorithm correctly classified 80.6% of subjects with incomplete KD who fulfilled AHA laboratory criteria. In contrast with the AHA algorithm, which requires an echocardiogram as part of the evaluation, this algorithm is intended for use at the point of care in settings where echocardiography would not be readily available. The algorithm correctly classified 80.6% of subjects with KD who manifested ≤3 KD principal criteria and were diagnosed by echocardiography. Furthermore, the algorithm correctly classified all 28 subjects with KD who went on to develop the most severe complication, coronary artery aneurysms.

There are both strengths and limitations to our study. We enrolled well-characterized, phenotypically similar control subjects, of whom approximately one-half were referred to our ED specifically for evaluation of possible KD. Thus, we used development and validation cohorts that mirror the patient population for which a classification algorithm would be most useful. In addition, the algorithm used widely available laboratory tests coupled with easily observable clinical signs and can be adapted as a computer- or smart phone-based tool. Nonetheless, 3.6% and 11.9% of subjects with KD and febrile control subjects, respectively, remained indeterminate. The algorithm performed less well when ≤3 clinical criteria were present. Although subject age did not adversely affect algorithm performance, illness day did have an impact with a greater proportion of febrile controls correctly identified early in the course of the illness. The algorithm had a higher sensitivity but lower specificity for subjects having 8 to 10 days of fever compared with those having ≤3 days of fever. The natural evolution of laboratory values in acute KD is for the inflammatory markers to diminish with time. Thus, by 8 to 10 days of fever, many of the key components of the algorithm were already starting to normalize, thus making some of the subjects with KD look more like the febrile controls.[13] The goal of KD management is to treat with intravenous immunoglobulin as soon as the diagnosis can be established. Thus, the fact that the algorithm performed well, discriminating patients with KD from febrile controls in the early phase of their illness, makes the

algorithm more valuable as a tool to ensure timely diagnosis and treatment. Integration into this algorithm of additional biomarkers that better differentiate subjects with KD and febrile controls could help to improve its performance. Incorporating nuanced clinical data such as limbal sparing of conjunctival injection or perineal accentuation of rash could also result in better diagnostic performance. Those data were not captured for this study, however, because the intention was to computationally capture the differences between patients to provide support for the more inexperienced practitioner. In the absence of a diagnostic test for KD diagnosis, there is always a possibility that subjects were classified erroneously. Thus, the best diagnostic tool will only be as good as expert clinicians until the etiology of KD is discovered and specific diagnostic tests can be devised. Before this algorithm can be adopted widely, it must be evaluated as a clinical device by the Food and Drug Administration, which will require prospective testing in larger cohorts from different medical centers where the "gold standard," including the use of echocardiography on febrile controls subjects, is established by other experts. The detailed algorithm will be made available to interested investigators upon request. ■

We thank our colleagues at the Stanford University Pediatric Proteomics Group for critical discussions, and the Stanford University IT group for Linux cluster support and assistance in data analysis and software development. We also thank Joan Pancheri, RN, BSN, for data collection and assistance with patient enrollment.

Submitted for publication Jan 8, 2016; last revision received Apr 14, 2016; accepted May 18, 2016.

Reprint requests: Xuefeng B. Ling, PhD, Department of Surgery, Stanford University, Stanford, CA 94305. E-mail: bxling@stanford.edu

References

1. Newburger JW, Takahashi M, Gerber MA, Gewitz MH, Tani LY, Burns JC, et al. Diagnosis, treatment, and long-term management of Kawasaki disease: a statement for health professionals from the Committee on Rheumatic Fever, Endocarditis and Kawasaki Disease, Council on Cardiovascular Disease in the Young, American Heart Association. Circulation 2004;110:2747-71.
2. Yellen ES, Gauvreau K, Takahashi M, Burns JC, Shulman S, Baker AL, et al. Performance of 2004 American Heart Association recommendations for treatment of Kawasaki disease. Pediatrics 2010;125:e234-41.
3. Rowley AH, Gonzalez-Crussi F, Gidding SS, Duffy CE, Shulman ST. Incomplete Kawasaki disease with coronary artery involvement. J Pediatr 1987;110:409-13.
4. Manlhiot C, Christie E, McCrindle BW, Rosenberg H, Chahal N, Yeung RS. Complete and incomplete Kawasaki disease: two sides of the same coin. Eur J Pediatr 2012;171:657-62.
5. Ha KS, Jang G, Lee J, Lee K, Hong Y, Son C, et al. Incomplete clinical manifestation as a risk factor for coronary artery abnormalities in Kawasaki disease: a meta-analysis. Eur J Pediatr 2013;172:343-9.
6. Ling XB, Kanegaye JT, Ji J, Peng S, Sato Y, Tremoulet A, et al. Point-of-care differentiation of Kawasaki disease from other febrile illnesses. J Pediatr 2013;162:183-8.e3.
7. Ling XB, Lau K, Kanegaye JT, Pan Z, Peng S, Ji J, et al. A diagnostic algorithm combining clinical and molecular data distinguishes Kawasaki disease from other febrile illnesses. BMC Med 2011;9:130.
8. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics 2001;17:520-5.
9. Oza NC, ed. Ensemble data mining methods. Hershey, PA: Idea Group Reference; 2006.
10. Breiman L. Random forests. Machine Learning 2001;45:5-32.
11. Breiman L. Bagging predictors. Machine Learning 1996;24:123-40.
12. Jaggi P, Kajon AE, Mejias A, Ramilo O, Leber A. Human adenovirus infection in Kawasaki disease: a confounding bystander? Clin Infect Dis 2013;56:58-64.
13. Tremoulet AH, Jain S, Chandrasekar D, Sun X, Sato Y, Burns JC. Evolution of laboratory values in patients with Kawasaki disease. Pediatr Infect Dis J 2011;30:1022-6.

Appendix 1

Additional members of the Pediatric Emergency Medicine Kawasaki Disease (KD) Research Group include (University of California San Diego, La Jolla, CA; Rady Children's Hospital San Diego, San Diego, CA): Lindsay T. Grubensky, RN, MSN, CPNP-PC, Jim R. Harley, MD, MPH, Paul Ishimine, MD, Jamie Lien, MD, Simon J. Lucio, MD, Seema Shah, MD, and Stacey Ulrich, MD.

Appendix 2

Methods

Missing Data Handling and Its Impact on Algorithm Performance. In our study cohort, there were 228 of 801 subjects with KD and 287 of 479 febrile controls with missing laboratory values. Of these, 340 (142 with KD, 198 febrile controls) were included in the development cohort, and 175 (86 with KD, 89 febrile controls) were included in the validation cohort.

Missing data were imputed using a method based on weighted K nearest neighbours algorithm as follows: Assume that a subject with KD S had missing values for variable v. The method would find other K ($K = 10$ in our study) subjects with KD that have complete value for v, and have most similarity to S in other $N - 1$ variables, where N is the total number of the clinical and laboratory test variables ($N = 18$) used in our study. The similarity was defined by the Euclidean distance between 2 subjects. Weighted average values of v in the K subjects with KD were used to estimate the missing value of v in the subject S. The weights of each of the K subjects were determined by its similarity to S. The same procedure was performed for febrile controls with missing values.

We explored the algorithm performance on subjects without and with missing laboratory values. For subjects without missing laboratory values, the primary LDA–based model correctly classified 91.1% of subjects (522 of 573) with KD and 44.3% of febrile controls (85 of 192), erroneously classified 13.0% of febrile controls (25 of 192), and 8.9% of subjects with KD (51 of 573); 42.7% of febrile controls (82 of 192) were left indeterminate. The secondary random forest models correctly classified 64.7% of subjects with KD (33 of 51) and 58.5% of febrile controls (48 of 82), erroneously classified 3.9% of subjects with KD (2 of 51) and 7.3% of febrile controls (6 of 82), and 31.4% of subjects with KD (16 of 51) and 34.1% of febrile controls (28 of 82) remained indeterminate. For subjects with missing laboratory values, the primary LDA-based model correctly classified 87.3% of subjects with KD (199 of 228) and 66.2% of febrile controls (190 of 287), erroneously classified 0.4% of subjects with KD (1 of 228) and 3.8% of febrile controls (11 of 287), and 12.3% of subjects with KD (28 of 228) and 30.0% of febrile controls (86 of 287) were left indeterminate. The secondary random forest models correctly classified 53.6% of subjects with KD (15 of 28) and 61.6% of febrile controls (53 of 86), erroneously classified 4.7% of febrile controls (4 of 86); 46.4% of subjects with KD (13 of 28) and 33.7% of febrile controls (29 of 86) remained indeterminate.

The overall performance of the 2-step algorithm on subjects without and with missing laboratory values is shown in Table VII. The combined results for subjects without missing values showed that 96.9% of subjects with KD (555 of 573) and 69.3% of febrile controls (133 of 192) were classified correctly, 0.3% of subjects with KD (2 of 573) and 16.1% of febrile controls (31 of 192) were classified erroneously, and 2.8% of subjects with KD (16 of 573) and 14.6% of febrile controls (28 of 192) were classified as indeterminate. The combined results for subjects with missing values showed that 93.9% of subjects with KD (214 of 228) and 84.7% of febrile controls (243 of 287) were classified correctly, 0.4% of subjects with KD (1 of 228) and 5.2% of febrile controls (15 of 287) were classified erroneously, and 5.7% of subjects with KD (13 of 228) and 10.1% of febrile controls (29 of 287) were classified as indeterminate.

Compared with the results shown in Figure 3, involvement of subjects with missing laboratory increased the indeterminate rate of subjects with KD from 2.8% to 3.6%, but reduced the indeterminate rate of febrile controls from 14.6% to 11.9%, leading to a reduced sensitivity (from 96.9 to 96.0%) but an increased specificity (from 69.3% to 78.5%). The positive predictive value (PPV) and negative predictive value (NPV) remained at similar levels (PPV decreased by 0.3% and NPV increased by 0.7%).

Appendix 3

Methods

Modeling Procedures of Random Forest Model. In modeling step I, subjects in the development cohort were further randomly partitioned into 2 subcohorts (subcohorts I and II) with equivalent sizes for model training [in step (I)] and calibration [in Step (II)]. Assuming there were N_1 subjects ($S_i, i = 1, 2, \ldots N_1$) with M clinical and laboratory test variables ($f_{ij}, j = 1, 2, \ldots M$) in subcohort I, each subject S_i in subcohort I can be expressed as $((f_i, y_i), i = 1, 2, \ldots N_1)$ where f_i is the M-dimensional vector for the subject S_i and y_i is the clinical outcome to be predicted (0 for KD and 1 for febrile controls). A "forest" of 300 forecasting decision trees was developed using the data in subcohort I. Specifically, each tree was developed using randomly selected 63.2% of the N_1 subjects and one-third of the M variables. At each node, the tree was split by choosing a split variable value producing the best split. The maximum size of each terminal node was 1. The final predicted score T was an average of decisions on each tree. That is, $T(S_i) = \frac{1}{300}\sum_{k=1}^{300} T_k,\ i = 1, \ldots N_1$.

In modeling step II, the predictive scoring threshold was calibrated on subcohort II to create a risk measure for an individual subject. Applying the step (I) model to each subject S_i in subcohort II, the derived predictive scores $T(S_i),\ i = 1, \ldots, N_2$ were ranked. For each value of T, we calculated the PPV and NPV for KD and febrile control classifications as follows:

$$PPV = f(T) = \sum_{i=1}^{N_2} I(T(S_i) - T)J(S_i) \Big/ \sum_{i=1}^{N_2} I(T(S_i) - T)$$

$$NPV = f(T) = \sum_{i=1}^{N_2} I(T - T(S_i))(1 - J(S_i)) \Big/ \sum_{i=1}^{N_2} I(T - T(S_i))$$

where. $I(x) = \left\{ \begin{array}{ll} 1 & x>0 \\ 0 & other \end{array} \right\}$ $J(x) = \left\{ \begin{array}{ll} 1 & x \in KD \\ 0 & x \in FC \end{array} \right\}$

In this way, we had a mathematic function mapping predictive values to PPV and NPV (ie, each subject was assigned a PPV and a NPV to estimate the risks of being a KD and febrile control with the given score). The 2 scores mapping to 0.95 PPV and NPV were selected as the cutoffs. We obtained 2 thresholds T_h and T_m from this mapping.

KD subgroup:

$$T(S_i) > T_h$$

Indeterminate subgroup:

$$T_m < T(S_i) < T_h$$

Febrile controls subgroup:

$$T(S_i) < T_m$$

In modeling step III, after calibration, the model's performance was tested using the validation cohort, assessing the model and calibration values derived in step (I) and (II). Again we applied the step (I) model to each subject S_i in the validation cohort to derive the predictive scores $T(S_i), i = 1, \ldots, N_3$ and determined the subgroup each subject belonged to according to the PPV and NPV score mapping constructed in step (II), so as to drive the decision for all subjects.

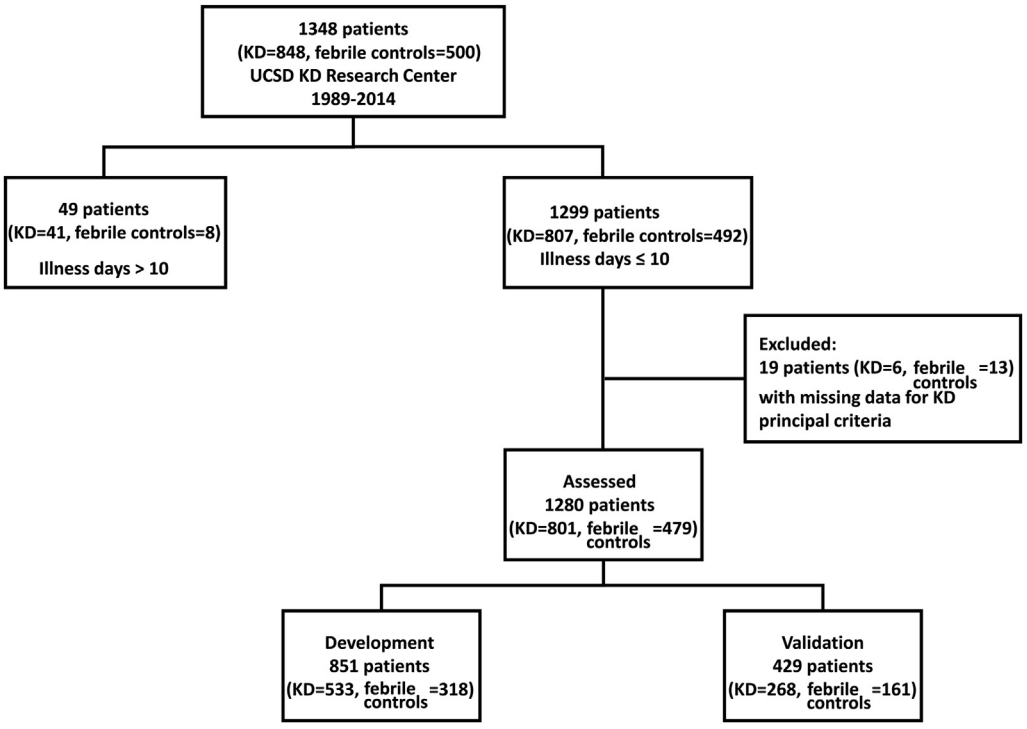


Figure 2. Cohort construction of retrospective development and validation.

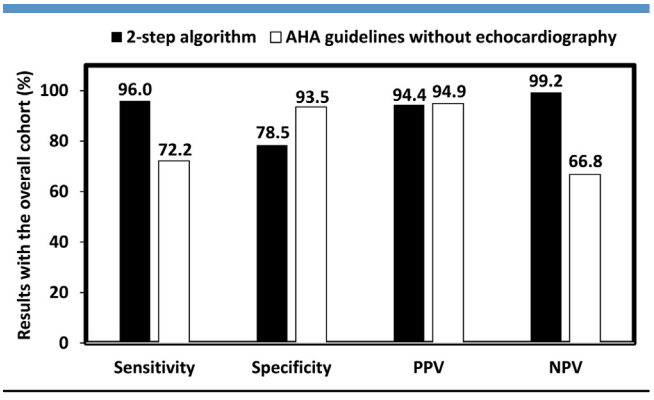


Figure 4. Comparison of subjects with KD and febrile controls differentiation performance on the overall cohort using the proposed 2-step algorithm and AHA guidelines at the absence of echocardiography, respectively.

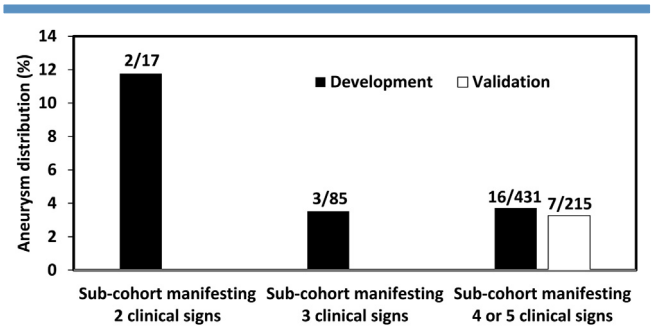


Figure 6. Distribution of the number of subjects who developed aneurysms over the 3 subcohorts in the development and validation cohorts, respectively.

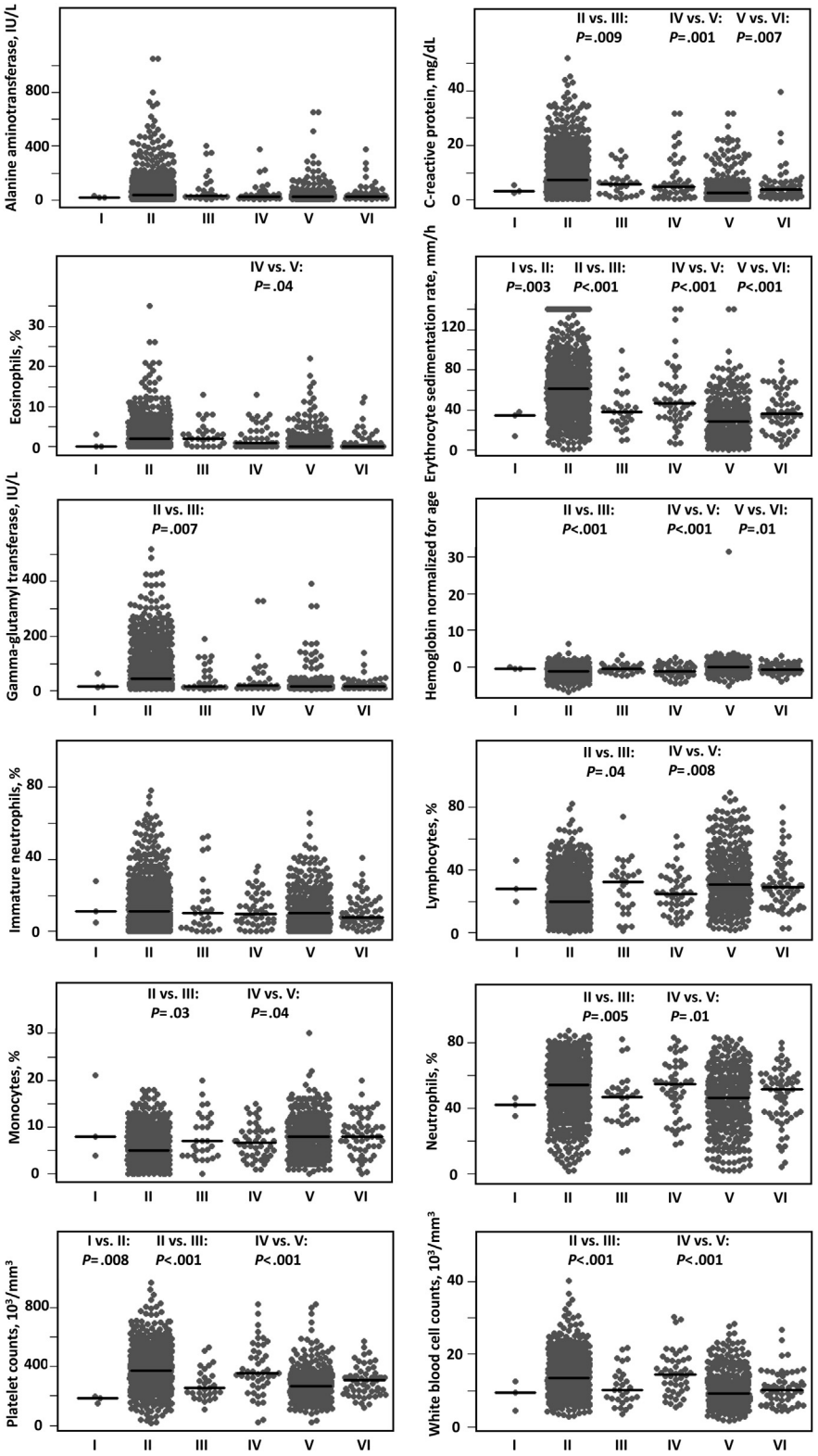


Figure 7. Distribution of each of the 12 laboratory test variables under all 6 classifications (I-VI) by 2-step algorithm. I: Erroneously classified KD; II: correctly classified KD; III: indeterminate KD; IV: erroneously classified febrile controls; V: correctly classified febrile controls; VI: indeterminate febrile controls. The median values of each variable were marked in black. Significant *P* values (<.05 with a Wilcoxon test) between the classifications are shown.

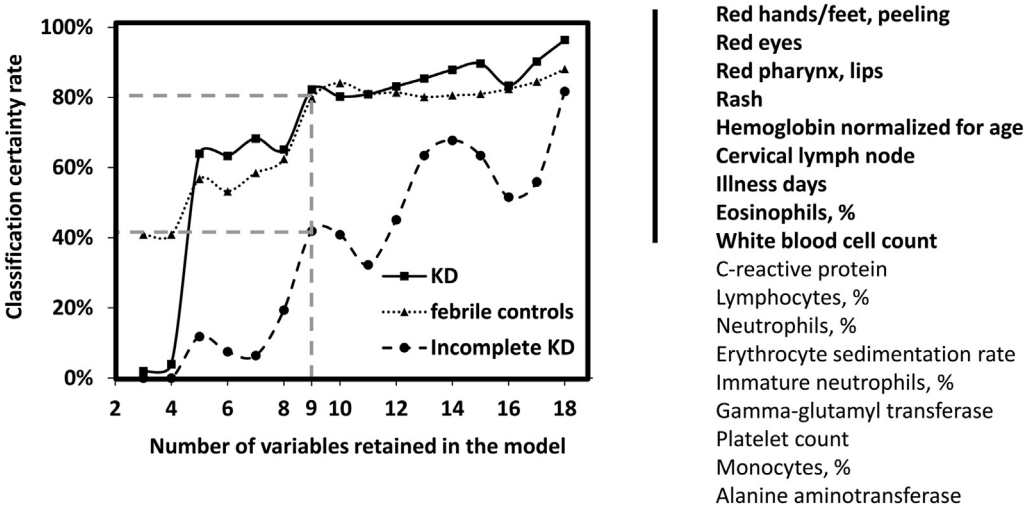


Figure 8. Performance of the 2-step algorithm developed with reduced number of variables. Proportions of KD, febrile controls, and incomplete subjects with KD with certain classification were calculated under each model. An 80% classification certainty rate was achieved with 9 variables: 5 KD principal clinical criteria, illness days, hemoglobin concentration normalized for age, percent eosinophils, and white blood cell count.

Table II. Final diagnoses of febrile controls in the development and validation cohorts

Diagnosis, n (%)	Development cohort (n = 318)	Validation cohort (n = 161)
Bacterial infections		
Methicillin-resistant *Staphylococcus aureus*	14 (4.4)	8 (5)
Scarlet fever*	14 (4.4)	4 (3)
Staphylococcal infection	13 (4.1)	2 (1)
Streptococcal pharyngitis	6 (1.9)	2 (1)
Cellulitis	6 (1.9)	0 (0)
Others	16 (5.0)	14 (9)
Viral infections		
Viral syndrome†	140 (44.0)	75 (47)
Adenovirus	43 (13.5)	28 (17)
Influenza virus	19 (6.0)	7 (4)
Enterovirus	14 (4.4)	0 (0)
Epstein–Barr virus	8 (2.5)	3 (2)
Others	5 (1.6)	4 (3)
Both bacterial and viral infection	3 (0.9)	2 (1)
Other/unknown‡	17 (5.3)	12 (8)

*Diagnosis of scarlet fever was based on a positive rapid streptococcal antigen test or throat culture and a compatible rash that responded to treatment with antibiotics active against this organism. Streptococcal serology was not consistently performed.
†Viral syndrome was defined as febrile illness without identified pathogen that spontaneously resolved without treatment.
‡Other/unknown included mycoplasma infection (3 in the development cohort, 2 in the validation cohort), drug reaction (5 in the development cohort, 1 in the validation cohort), and unknown diagnosis (9 in the development cohort, 9 in the validation cohort).

Table III. Distribution of the clinical signs among subjects in subcohorts of the development cohort manifesting 1, 2, 3, or ≥4 clinical criteria for KD

	Clinical criteria								
	1	2			3			≥4		
Clinical signs, n (%)	Febrile controls (n = 126)	KD (n = 17)	Febrile controls (n = 108)	*P**	KD (n = 85)	Febrile controls (n = 56)	*P**	KD (n = 431)	Febrile controls (n = 28)	*P**
Cervical lymph node	14 (11)	3 (18)	12 (11)	.43	10 (12)	21 (38)	<.001	173 (40.1)	13 (46.4)	.55
Rash	91 (72)	12 (71)	86 (80)	.53	71 (84)	47 (84)	.99	420 (97.4)	28 (100)	.99
Red eyes	12 (10)	37 (34)	8 (47)	.42	66 (78)	45 (80)	.83	416 (96.6)	27 (96)	.99
Red hands/feet, peeling	1 (1)	1 (6)	25 (23)	.19	43 (51)	12 (21)	<.001	400 (92.8)	20 (71)	.001
Red pharynx, lips	8 (6)	10 (59)	56 (52)	.61	65 (77)	43 (77)	.99	422 (97.9)	28 (100)	.99

*Fisher exact test.

Table IV. Multivariate analysis of illness duration and laboratory test variables in discriminating KD from febrile control subjects in subcohorts of the development cohort with 2, 3, or ≥4 clinical criteria for KD

	2 clinical criteria		3 clinical criteria		≥4 clinical criteria	
Variables	OR (95% CI)	*P**	OR (95% CI)	*P**	OR (95% CI)	*P**
Illness days	1.87 (1.00-3.54)	.04	1.19 (0.91-1.56)	.20	1.31 (1.02-1.70)	.03
Alanine aminotransferase	0.99 (0.95-1.03)	.50	1.01 (1.00-1.02)	.07	1.00 (1.00-1.01)	.22
CRP	1.00 (0.84-1.20)	.97	1.14 (1.01-1.29)	.02	1.06 (0.97-1.15)	.15
Eosinophils, %	0.91 (0.57-1.44)	.66	1.12 (0.97-1.30)	.13	1.07 (0.90-1.27)	.46
Erythrocyte sedimentation rate	1.03 (0.98-1.08)	.26	1.04 (1.01-1.07)	.01	1.03 (1.01-1.05)	.003
Gamma-glutamyl transferase	1.04 (1.01-1.08)	.01	1.01 (0.99-1.02)	.29	1.01 (1.00-1.02)	.12
Hemoglobin normalized for age	0.49 (0.20-1.21)	.11	0.78 (0.52-1.17)	.21	0.84 (0.63-1.12)	.23
Immature neutrophils, %	0.90 (0.73-1.12)	.28	0.98 (0.91-1.07)	.68	0.97 (0.92-1.02)	.22
Lymphocytes, %	1.15 (0.92-1.43)	.22	0.98 (0.91-1.05)	.51	0.97 (0.91-1.02)	.22
Monocytes, %	1.08 (0.77-1.53)	.65	0.96 (0.83-1.11)	.57	0.96 (0.83-1.11)	.55
Neutrophils, %	1.18 (0.96-1.45)	.10	0.99 (0.92-1.07)	.85	0.95 (0.90-1.01)	.07
Platelet count	1.01 (1.00-1.02)	.25	1.01 (1.00-1.01)	.03	1.00 (1.00-1.01)	.74
White blood cell count	1.46 (1.10-1.96)	.003	0.98 (0.86-1.11)	.71	1.05 (0.94-1.18)	.33

*Likelihood ratio test.

Table V. Variable importance* in the secondary random forest model for subcohorts of subjects manifesting 2, 3, or ≥ 4 clinical criteria for KD†

Variable	2 clinical criteria	3 clinical criteria	≥4 clinical criteria
Cervical lymph node	−1.00	2.45	0.00
Rash	0.00	−0.73	0.00
Red eyes	0.82	−1.38	−1.00
Red hands/feet, peeling	2.99	2.99	0.73
Red pharynx, lips	−0.58	1.35	0.00
Illness days	2.46	1.61	1.80
Alanine aminotransferase	−3.63	2.58	5.83
CRP	1.59	6.04	−1.05
Eosinophils, %	−0.19	3.56	−1.15
Erythrocyte sedimentation rate	5.58	7.79	1.53
Gamma-glutamyl transferase	−1.45	1.67	1.14
Hemoglobin normalized for age	0.83	9.16	0.32
Immature neutrophils, %	0.15	7.21	0.71
Lymphocytes, %	1.92	0.52	1.33
Monocytes, %	2.19	−0.26	−1.61
Neutrophils, %	5.76	2.76	−0.37
Platelet count	1.83	8.25	0.76
White blood cell count	6.55	0.42	2.66

*Measured by percent increase of model mean square error caused by permutation of variable values.
†The higher the number, the greater the importance of the variable in discriminating KD from febrile controls for each of the subcohorts based on number of clinical criteria.

Table VI. Performance of the algorithm on subcohorts stratified by age at onset and illness day

Subcohorts	Sensitivity, % (95% CI)	Specificity, % (95% CI)	PPV, % (95% CI)	NPV, % (95% CI)	Indeterminate KD, %	Indeterminate febrile controls, %
Age at onset*						
≤6 months (KD, n = 69; febrile controls, n = 23)	97.1 (89.9-99.6)	87.0 (66.4-97.2)	100 (92.1-100)	100 (76.2-100)	2.9	13.0
>6 months (KD, n = 732; febrile controls, n = 456)	95.9 (94.2-97.2)	78.1 (74.0-81.8)	93.9 (91.9-95.5)	99.2 (97.6-99.8)	3.7	11.8
Illness days						
≤3 (KD, n = 83; febrile controls, n = 168)	95.2 (88.1-98.7)	85.7 (79.5-90.6)	89.8 (81.5-91.2)	99.3 (96.2-100)	3.6	8.9
4-5 (KD, n = 294; febrile controls, n = 141)	94.2 (90.9-96.6)	78.0 (70.3-84.5)	96.2 (93.3-98.1)	98.2 (93.7-99.8)	5.1	14.2
6-7 (KD, n = 269; febrile controls, n = 121)	97.0 (94.2-98.7)	76.0 (67.4-83.3)	95.3 (92.0-97.4)	100 (94.2-100)	3.0	13.2
8-10 (KD, n = 155; febrile controls, n = 49)	98.1 (94.4-99.6)	61.2 (46.2-74.8)	92.1 (86.9-95.7)	100 (83.3-100)	1.9	12.2

*Age at the first day of fever.

Table VII. Performance of the algorithm on subjects with no missing and missing laboratory values

Cohort	Sensitivity, % (95% CI)	Specificity, % (95% CI)	PPV, % (95% CI)	NPV, % (95% CI)	Indeterminate KD, %*	Indeterminate febrile controls, %*
No missing						
Development (KD, n = 391; febrile controls, n = 120)	97.7 (95.7-98.9)	75.8 (67.2-83.2)	95.5 (93.0-97.3)	1 (94.1-1)	2.3	9.2
Validation (KD, n = 182; febrile controls, n = 72)	95.1 (90.8-97.7)	58.3 (46.1-69.8)	93.0 (88.3-96.2)	95.5 (84.5-99.4)	3.8	23.6
Combined (KD, n = 573; febrile controls, n = 192)	96.9 (95.1-98.1)	69.3 (62.2-75.7)	94.7 (92.6-96.4)	98.5 (94.8-99.8)	2.8	14.6
Missing						
Development (KD, n = 142; febrile controls, n = 198)	93.7 (88.3-97.1)	86.4 (80.8-90.8)	92.4 (86.7-96.1)	99.4 (96.8-1)	5.6	8.1
Validation (KD, n = 86; febrile controls, n = 89)	94 (87-98)	81 (71-88)	95 (84-99)	1 (93-1)	6	15
Combined (KD, n = 228; febrile controls, n = 287)	93.9 (89.9-96.6)	84.7 (80.0-88.6)	93.4 (89.4-96.3)	99.6 (97.7-1)	5.7	10.1

*The percent of subjects with KD or febrile control subjects classified as indeterminate by the algorithm.