


# Automated Identification and Extraction of Exercise Treadmill Test Results

Chengyi Zheng , PhD, MS; Benjamin C. Sun, MD, MPP; Yi-Lin Wu, MS; Ming-Sum Lee, MD, PhD; Ernest Shen, PhD; Rita F. Redberg, MD, MSc; Maros Ferencik, MD, PhD, MCR; Shaw Natsui, MD, MPA; Aniket A. Kawatkar, PhD, MS; Visanee V. Musigdilok, MPH; Adam L. Sharp, MD, MS

**Background**—Noninvasive cardiac tests, including exercise treadmill tests (ETTs), are commonly utilized in the evaluation of patients in the emergency department with suspected acute coronary syndrome. However, there are ongoing debates on their clinical utility and cost-effectiveness. It is important to be able to use ETT results for research, but manual review is prohibitively time-consuming for large studies. We developed and validated an automated method to interpret ETT results from electronic health records. To demonstrate the algorithm's utility, we tested the associations between ETT results with 30-day patient outcomes in a large population.

**Methods and Results**—A retrospective analysis of adult emergency department encounters resulting in an ETT within 30 days was performed. A set of randomly selected reports were double-blind reviewed by 2 physicians to validate a natural language processing algorithm designed to categorize ETT results into normal, ischemic, nondiagnostic, and equivocal categories. Natural language processing then searched and categorized results of 5214 ETT reports. The natural language processing algorithm achieved 96.4% sensitivity and 94.8% specificity in identifying normal versus all other categories. The rates of 30-day death or acute myocardial infarction varied ( $P < 0.001$ ) by categories for normal (0.08%), ischemic (1.9%), nondiagnostic (0.77%), and equivocal (0.58%) groups achieving good discrimination (C-statistic, 0.81; 95% CI, 0.7–0.92).

**Conclusions**—Natural language processing is an accurate and efficient strategy to facilitate large-scale outcome studies of noninvasive cardiac tests. We found that most patients are at low risk and have normal ETT results, while those with abnormal, nondiagnostic, or equivocal results have slightly higher risks and warrant future investigation. (*J Am Heart Assoc.* 2020;9:e014940. DOI: 10.1161/JAHA.119.014940.)

**Key Words:** cardiac event • chest pain • emergency department • natural language processing • noninvasive test • treadmill test

Noninvasive cardiac tests, including exercise treadmill tests (ETTs), are recommended in the evaluation of patients with suspected acute coronary syndrome.<sup>1,2</sup> However, the benefits of routine use of noninvasive cardiac tests remains unclear as there is no evidence for reduction in death or acute myocardial infarction (AMI).<sup>3–5</sup> Because of the costs and risks associated with noninvasive test strategies,<sup>6,7</sup> there

is a strong need for comparative effectiveness studies to assess the value of ETT in acute care settings.<sup>3,4</sup>

An essential technical barrier to such studies is the need to extract clinical information from ETT text reports. Because of low event rates and confounding factors in observational data, an adequately powered study would require clinical data from vast numbers of ETTs.<sup>8</sup> With  $\approx 1$  million ETTs performed since

From the Research and Evaluation Department, Kaiser Permanente Southern California, Pasadena, CA (C.Z., Y.-L.W., E.S., A.A.K., V.V.M., A.L.S.); Department of Emergency Medicine, University of Pennsylvania, Philadelphia, PA (B.C.S.); Division of Cardiology, Kaiser Permanente Southern California, Los Angeles Medical Center, Los Angeles, CA (M.-S.L.); Division of Cardiology, University of California, San Francisco, San Francisco, CA (R.F.R.); Knight Cardiovascular Institute, Oregon Health and Science University, Portland, OR (M.F.); National Clinician Scholars Program, Department of Emergency Medicine, University of California, Los Angeles, Los Angeles, CA (S.N.).

Accompanying Data S1 through S7 and Tables S1 through S5 are available at <https://www.ahajournals.org/doi/suppl/10.1161/JAHA.119.014940>

**Correspondence to:** Chengyi Zheng, PhD, MS, Department of Research and Evaluation, Kaiser Permanente Southern California, 100 South Los Robles Avenue, 2nd Floor, Pasadena, CA 91101. E-mail: [chengyi.x.zheng@kp.org](mailto:chengyi.x.zheng@kp.org)

Received October 14, 2019; accepted January 30, 2020.

© 2020 The Authors. Published on behalf of the American Heart Association, Inc., by Wiley. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

## Clinical Perspective

### What Is New?

- Exercise treadmill test (ETT) reports have a rich set of information with diagnostic and prognosis value but are challenging to use because of their unstructured format.
- Natural language processing provides an efficient way to identify and extract ETT variables from ETT reports.
- The majority of patients in the emergency department who underwent ETT had normal results and were at low risk, and patients with inconclusive ETT results (equivocal and nondiagnostic) were significantly different.

### What Are the Clinical Implications?

- This study demonstrates that ETT shows good prediction on near-term cardiac outcomes.
- ETT may offer a better value proposition as a prognostic tool compared with a diagnostic tool.
- Instead of treating equivocal and nondiagnostic as inconclusive ETT tests, as is commonly done in current clinical practice, these patients may warrant different treatment pathways.

2000 in our regional health system alone, there is tremendous interest in using the information documented in these test reports for research. However, clinical ETT data are typically in a free-text format. Studies have required manual review of noninvasive test results, which is time-consuming and expensive. An automated method that can extract information documented in the unstructured testing reports would greatly facilitate studies that require data from large numbers of ETT reports.

With the widespread use of electronic health record (EHR) systems, clinical notes are electronically available. Natural language processing (NLP) is a computer-based method that has been utilized to identify and extract information from clinical notes. When compared with manual chart review of medical records, NLP is more efficient and produces more consistent results.<sup>9,10</sup> Our team has previously developed NLP algorithms for cardiovascular variables, such as extraction of ejection fraction from echocardiography reports.<sup>11–13</sup> The goals of this study were to: (1) derive and validate an algorithm to identify ETT results from unstructured reports, and (2) demonstrate the algorithm's utility by correlating ETT results with 30-day patient outcomes in a large population.

## Methods

The data, analytic methods, and study materials will not be made available to other researchers for purposes of reproducing the results or replicating the procedure.

## Study Setting

This retrospective cohort study was conducted at Kaiser Permanente Southern California (KPSC), an integrated health-care organization with over 7600 physicians, 15 medical centers, and 231 medical offices. KPSC provides prepaid comprehensive health care to 4.6 million racially and socioeconomically diverse members. Members receive medical care in KPSC-owned facilities and contracting facilities. All KPSC emergency department (ED) sites use the same troponin laboratory assay (Beckman Coulter Access AccuTnI+3) with an AMI threshold level of 0.5 ng/mL, and ED physicians can order noninvasive cardiac testing as part of the discharge and follow-up plan of patients with suspected acute coronary syndrome.

## Study Population

We included all KPSC members 18 years or older with an ED visit between January 1, 2015, to September 19, 2017, and who had a troponin laboratory test and underwent an ETT within 30 days of their ED visits. We excluded patients who were transferred from a non-KPSC hospital or died during ED visits. We excluded patients without KPSC health plan membership during the 12 months before and 30 days after ED visits because accurate comorbidities and patient outcomes are not available for nonmembers. Noninvasive cardiac tests were identified by Current Procedural Terminology codes (Data S1).

Patient demographic information such as age, sex, and race were obtained from administrative records. HEART (history, ECG, age, risk factors, and troponin) is a risk score used to inform clinical decision making<sup>14</sup> and KPSC implemented the HEART score into routine ED care in May 2016.<sup>15</sup> Therefore, HEART scores calculated at the time of the index ED visit were captured in the EHR when available, as well as other variables such as smoking history. As in previous reports, *International Classification of Diseases, Ninth and Tenth Revision (ICD-9 and ICD-10)* codes in the structured EHR data were used to define coronary artery disease, diabetes mellitus, dyslipidemia, hypertension, stroke, and the Elixhauser comorbidity index.<sup>16,17</sup>

## Training and Validation Data Sets

Based on the sample size calculation,<sup>18</sup> using a prevalence rate of non-normal findings among ETT of 32% (32%,<sup>19</sup> 36%,<sup>20</sup> and 39%<sup>21</sup> in previous studies), the minimal size of the validation data set is 84 when the expected precision of estimate (ie, the maximum marginal error) is 0.1 and CI is 95%. Therefore, among the study population, we performed random sampling to create NLP training (n=115) and validation (n=115) data sets. Ten patients were excluded from the validation data because there were no associated ETT reports.

The ETT reports of the remaining 105 patients in the validation data set were reviewed independently by an emergency physician (A.L.S.) and a cardiologist (M.S.L.). Besides the final ETT impression, the physician review also abstracted additional information from the ETT reports (Data S2). ETT reports were primarily to assist reviewers and the NLP algorithm to appropriately categorize patients into ischemic, nondiagnostic, equivocal, or normal categories. The following are the simplified definitions for each category:

**Ischemic:** Cardiologist-reported ischemic changes or abnormal ST results defined as an upsloping ST change  $\geq 2$  mm or downsloping or horizontal ST change  $\geq 1$  mm.

**Nondiagnostic:** Patient heart rate (HR) does not rise to 85% of the maximum predicted HR during ETT.

**Equivocal:** Any abnormal results that were not categorized by ischemic or nondiagnostic definitions.

**Normal:** Patient completed the ETT with an appropriate maximum predicted HR and no ischemic ECG changes or other significant abnormalities.

Other definitions used to categorize ETT results are found in Data S3. The results of physician review were compared, and discrepancies were resolved by consensus and discussion with the other physicians on the research team (B.S., M.F., R.F.R.). The adjudicated results served as the reference standard against which NLP was compared.

## NLP Algorithm Development

The NLP modules used in this study were previously described.<sup>9,11</sup> Terminologies were created to capture ETT-related information (Data S4). The NLP search was performed for each report on 3 levels: sentence, neighboring sentences, and section (Data S5). A relationship detection algorithm was applied to relate the identified symptoms to the corresponding time periods. Negation and temporal relationship detection algorithms were applied to identify and exclude negated, uncertain, historical, and future statements. Negation algorithm handles double negations that commonly occur in ETT reports, eg, “no significant abnormality.” Regular expressions were created to capture some of the values. We developed separated algorithms to identify and extract each clinical variable commonly available in ETT reports (Figure 1, Data S2, and Table S1).

A postprocessing step was developed using Python programming language to integrate and finalize the results. Additional variables were derived based on the NLP-extracted variables and the variables (age and sex) from structured EHR data (Data S2 and Table S2). A data imputing step was performed to fill missing data using other variables. For example, based on the age and maximum HR, maximum predicted HR can be calculated (Data S3). Based on the

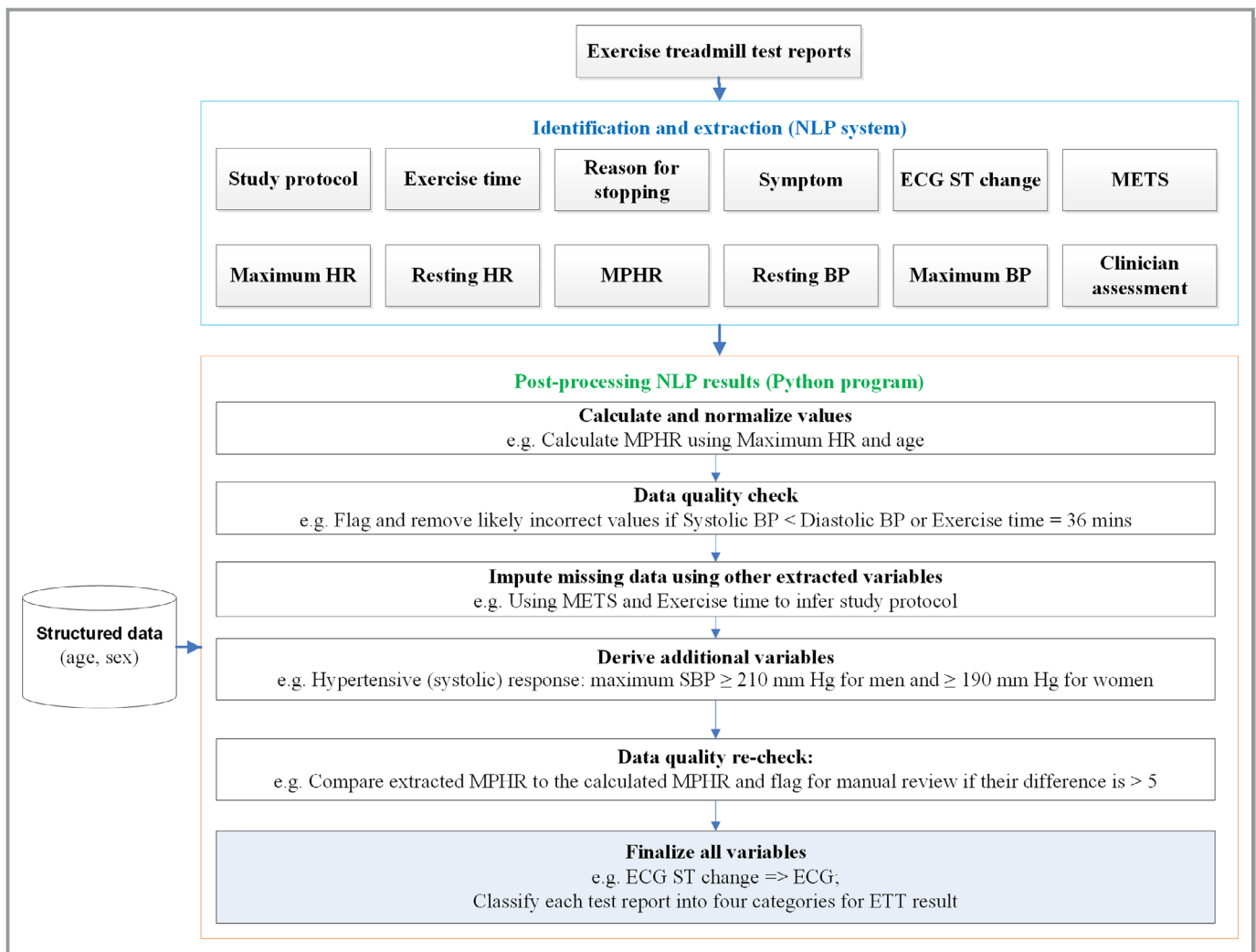
exercise time and metabolic equivalents (METs), it can infer whether it is the standard Bruce protocol (Data S6). Algorithms have also been developed to identify incorrect information in the reports. For example, incorrect values were flagged and discarded if they were out of the clinical range, such as an MET of 50. The magnitude of ST change and its direction was used to classify the ECG result into normal, abnormal, and equivocal categories (Data S7).<sup>22</sup> The ETT results were classified as abnormal, normal, equivocal, and nondiagnostic categories based on the clinician’s assessment as well as the other information documented in the reports (Data S3).<sup>22,23</sup> The NLP algorithm was developed and iteratively improved using the training data set.

## NLP Algorithm Validation

The performance of NLP was evaluated against the validation data set at the patient level. Confusion matrix, a type of classification table commonly used in the visualization of the performance of a machine learning classification algorithm, was depicted to compare the NLP results to the reference standard for identification of ETT results. The multicategory variables were dichotomized into 2 categories for evaluation purposes. The numbers of true positives, false positives, true negatives, and false negatives were calculated for each variable. Sensitivity, specificity, positive predictive value, negative predictive value, and negative/positive likelihood ratios were then derived based on those numbers.

## Application of NLP Algorithm and Analysis

NLP algorithms were further refined based on the validation results. The final NLP algorithm was then applied to the entire study population of patients with exercise testing to identify the ETT results. Patient characteristics and comorbidities were compared among the different ETT results. The ETT result was treated as a nominal variable rather than an ordinal variable. The primary outcome was 30-day AMI or all-cause mortality. The secondary outcome was 30-day major adverse cardiac event rates, which was the composite of death, AMI, and any coronary revascularization procedures. We calculated *P* values using chi-square or Fisher exact tests for all categorical variables and Wilcoxon test for all continuous variables. The significance threshold was set at 0.05. To reduce potential bias for rare events, logistic regression with Firth penalized maximum likelihood method<sup>24</sup> was used to estimate odds ratios (ORs) and 95% CIs. C-statistics were calculated for the ETT’s ability to predict the primary and secondary outcomes. All data were analyzed using SAS version 9.4 (SAS Institute Inc.). The institutional review board at KPSC approved this study. Requirement for informed consent was waived.



**Figure 1.** Diagram illustrating the natural language processing (NLP) process to extract and process exercise treadmill test (ETT) reports. BP indicates blood pressure; HR, heart rate; METs, metabolic equivalents; MPHR, maximum predicted heart rate; SBP, systolic blood pressure.

## Results

Our study population included 5214 patients with a median age of 56 years, 50.4% were women, and 48.1% were white (Table 1). The interannotator agreements (Cohen  $\kappa$ ) on the validation data set were reported in Table S3. The overall agreements are substantial to excellent based on Landis and Koch.<sup>25</sup> In the reference standard, the percentages of abnormal, equivocal, nondiagnostic, and normal ETT results were 5.7%, 6.7%, 14.3%, and 73.3%, respectively. NLP achieved 96.4% sensitivity and 94.8% specificity on identifying non-normal (abnormal/equivocal/nondiagnostic) versus normal ETT tests (Table 2) on the validation data set. The positive predictive value was 87.1% and the negative predictive value was 98.6%. NLP had the highest accuracy in identifying nondiagnostic results. For abnormal and equivocal results, NLP had higher specificity and negative predictive value but lower sensitivity and positive predictive value. The evaluation

results for the other 9 ETT variables are presented in Table 3. NLP achieved high accuracy on these variables except for the relatively low positive predictive value for symptom identification.

The refined NLP algorithm was applied to the 5214 ETT reports. The percentages of abnormal, equivocal, nondiagnostic, and normal ETT results were 5.9%, 6.6%, 12.5%, and 75%, respectively. Table 1 shows patient characteristics stratified by the ETT results. The troponin values were reported in Table S4. Most of these patients had a troponin value  $<0.02$  ng/mL.

The mean and median days from ED to ETT were 4 and 1, respectively. Bruce protocol was used in 95% of patients. Table 4 presents the ETT variables stratified by the ETT results. Compared with the patients with normal ETT results, the other groups were more likely to have shorter exercise time, lower METs, lower maximum HR, and chronotropic incompetence.

**Table 1.** Comparison of Patient Characteristics by Treadmill Test Results

Patient Variables	Normal	Abnormal	Equivocal	Nondiagnostic	P Value*	Total
No. (%)	3908 (75)	310 (5.9)	344 (6.6)	652 (12.5)		5214 (100)
Age, y <sup>†</sup>	55 (47, 64)	58 (50, 65)	57 (49, 64)	60 (52, 69)	<0.001	56 (48, 65)
Women	1955 (50)	138 (44.5)	182 (52.9)	355 (54.4)	0.022	2630 (50.4)
Hispanic	1591 (40.7)	123 (39.7)	129 (37.5)	278 (42.6)	0.68	2121 (40.7)
Race					0.32	
White	1895 (48.5)	154 (49.7)	166 (48.3)	294 (45.1)		2509 (48.1)
Black	400 (10.2)	37 (11.9)	42 (12.2)	90 (13.8)		569 (10.9)
Asian	492 (12.6)	42 (13.5)	47 (13.7)	86 (13.2)		667 (12.8)
Alaska Native/Pacific Islander	79 (2)	3 (1)	6 (1.7)	9 (1.4)		97 (1.9)
Other	1042 (26.7)	74 (23.9)	83 (24.1)	173 (26.5)		1372 (26.3)
Smoking behavior					0.003	
Never	2548 (65.2)	203 (65.5)	240 (69.8)	393 (60.3)		3384 (64.9)
Other	1253 (32.1)	100 (32.3)	102 (29.7)	249 (38.2)		1704 (32.7)
HEART score	3 (2, 4)	4 (3, 4)	3 (2, 4)	4 (2, 5)	0.009	1065 (20.4)
HEART score (risk groups)					0.12	
Low (0–3)	468 (58.6)	32 (46.4)	44 (60.3)	60 (48)		604 (56.7)
Intermediate (4–6)	320 (40.1)	35 (50.7)	27 (37)	63 (50.4)		445 (41.8)
High (≥7)	10 (1.3)	2 (2.9)	2 (2.7)	2 (1.6)		16 (1.5)
Elixhauser index	2 (1, 3)	2 (1, 4)	2 (1, 4)	3 (2, 5)	<0.001	5214 (100)
Comorbidities						
Coronary artery disease	217 (5.6)	51 (16.5)	29 (8.4)	95 (14.6)	<0.001	392 (7.5)
Stroke	31 (0.8)	4 (1.3)	2 (0.6)	11 (1.7)	0.12	48 (0.9)
Dyslipidemia	2279 (58.3)	203 (65.5)	206 (59.9)	437 (67)	<0.001	3125 (59.9)
Hypertension	1605 (41.1)	179 (57.7)	166 (48.3)	419 (64.3)	<0.001	2369 (45.4)
Diabetes mellitus	756 (19.3)	96 (31)	76 (22.1)	210 (32.2)	<0.001	1138 (21.8)
Medications, No. (%) <sup>‡</sup>						
Anticoagulants	109 (2.8)	15 (4.8)	18 (5.2)	52 (8)	<0.0001	194 (3.7)
Hyperlipidemics	965 (24.7)	104 (33.5)	98 (28.5)	247 (37.9)	<0.0001	1414 (27.1)
Hypertensives	1233 (31.6)	139 (44.8)	122 (35.5)	351 (53.8)	<0.0001	1845 (35.4)
Diabetes mellitus	421 (10.8)	58 (18.7)	49 (14.2)	134 (20.6)	<0.0001	662 (12.7)

HEART indicates history, ECG, age, risk factors, and troponin.

\*Chi-square test was used for categorical variables, and Wilcoxon test was used for continuous variables.

<sup>†</sup>Continuous variables are expressed as median (25th, 75th percentiles). Data are presented as number (percentage) unless otherwise indicated.

<sup>‡</sup>Medication usage in the 90 days before emergency department visits.

Overall event rates were low (Table 5, Figure 2). There were associations of increasing 30-day death/AMI with ETT results ( $P<0.001$ ) from normal (0.08%; 95% CI, 0–0.16), to nondiagnostic (0.77%; 95% CI, 0.1–1.44), to equivocal (0.58%; 95% CI, 0–1.38), to abnormal (1.9%; 95% CI, 0.4–3.47). There were stronger associations of increasing 30-day major adverse cardiac event rates with ETT results ( $P<0.001$ ) from normal (0.08%; 95% CI, 0–0.16), to nondiagnostic (1.1%; 95% CI, 0.28–1.86), to equivocal (2.03%; 95% CI, 0.54–3.53), to abnormal (10.0%; 95% CI, 6.66–13.34).

Table 6 presents the unadjusted ORs for ETT results in patients who had 30-day major adverse cardiac event rates or death/AMI versus patients who did not. Compared with normal ETT, nondiagnostic, equivocal, and abnormal ETT were associated with higher odds of 30-day death/AMI (nondiagnostic: OR, 9.5 [95% CI, 2.5–40.9]; equivocal: OR, 8.1 [95% CI, 1.4–42.0]; and abnormal: OR, 23.8 [95% CI, 6.7–100.4]). The C-statistic was 0.81 (95% CI, 0.70–0.92). Compared with normal ETT, nondiagnostic, equivocal, and abnormal ETT were associated with higher odds of 30-day major adverse cardiac

**Table 2.** Comparison of NLP to the Reference Standard for Identification of ETT Results

Confusion Matrix	NLP				Total
	Normal	Abnormal	Equivocal	Nondiagnostic	
Normal	73	1	3		77
Abnormal	1	5			6
Equivocal		2	7		9
Nondiagnostic				13	13
Total	74	8	10	13	
Comparison Groups*	Sensitivity, % (95% CI)		Specificity, % (95% CI)		NPV, % (95% CI)
Normal vs rest	96.4 (79.8–99.8)		94.8 (86.5–98.3)		98.6 (91.7–99.9)
Abnormal vs rest	83.3 (36.5–99.1)		97.0 (90.8–99.2)		99.0 (93.6–99.9)
Equivocal vs rest	77.8 (40.2–96.1)		96.9 (90.5–99.2)		97.9 (91.9–99.6)
Nondiagnostic vs rest	100 (71.7–100)		100 (95.0–100)		100 (95.0–100)

NLP indicates natural language processing; NPV, negative predictive value; PPV, positive predictive value.

\*For evaluation purposes, the multicategory exercise treadmill test (ETT) results were dichotomized into 2 categories.

event rates (nondiagnostic: OR, 13 [95% CI, 3.8–53.5]; equivocal: OR, 24.8 [95% CI, 7.3–102.5]; and abnormal: OR, 125.8 [95% CI, 47.2–466.3]). The C-statistic was 0.9 (95% CI, 0.86–0.95).

## Discussion

In the era of big data, unstructured (or free-text) data in the EHR has become an increasingly valuable source for clinical research and operational measurement. However, the

traditional approach of using unstructured data requires manual chart review. Manual chart review is not only time-consuming and costly but it often lacks accuracy and consistency.<sup>26</sup> In this study, we derived and validated a highly accurate automatic algorithm using NLP to identify, extract, and synthesize information from free-text ETT reports. The NLP algorithm had high sensitivity and specificity compared with physician reviewers and accurately identified normal, ischemic, nondiagnostic, and equivocal ETT results. We expect these results would yield similar results in different systems as we have found previous NLP

**Table 3.** Comparison of NLP to the Reference Standard for Identification of Treadmill Test Variables

ETT Variables	Reference Standard (n/N)	Sensitivity % (95% CI)	Specificity % (95% CI)	PPV % (95% CI)	NPV % (95% CI)
Study protocol*	98/105	95.9 (89.3–98.7)	100 (77.1–100)	100 (95.1–100)	81 (57.4–93.7)
Exercise time	104/105	94.2 (87.4–97.6)	100 (67.9–100)	100 (95.3–100)	64.7 (38.6–84.7)
Reasons for stopping*	92/105	98.9 (93.2–99.9)	100 (82.2–100)	100 (95–100)	95.8 (76.9–99.8)
Symptom*	100/105	80 (29.9–98.9)	94 (86.9–97.5)	40 (13.7–72.6)	98.9 (93.4–99.9)
Symptom2*	89/105	100 (39.6–100)	98.8 (92.7–99.9)	80 (29.9–98.9)	100 (94.6–100)
ECG*	105/105	98.1 (92.6–99.7)	100 (67.9–100)	100 (95.5–100)	84.6 (53.7–97.3)
METs	104/105	100 (95.6–100)	100 (67.9–100)	100 (95.6–100)	100 (67.9–100)
Maximum BP	96/105	96.9 (90.5–99.2)	100 (79.1–100)	100 (95.1–100)	86.4 (64–96.4)
MpHR	104/105	100 (95.6–100)	100 (67.9–100)	100 (95.6–100)	100 (67.9–100)
Maximum HR	94/105	90.4 (82.2–95.3)	100 (80.8–100)	100 (94.6–100)	70 (50.4–84.6)

The results of sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) findings were reported as percentages with 95% CIs. BP indicates blood pressure; ETT, exercise treadmill test; METs, metabolic equivalents; MpHR, maximum predicted heart rate; NLP, natural language processing.

\*For evaluation purposes, the results of these multicategory variables were dichotomized into 2 categories:

1. Study protocol: standard Bruce protocol vs other types of study protocols.
2. Reasons for stopping: target heart rate (HR) achieved vs other reasons.
3. Symptom: no symptoms vs abnormal, atypical angina, atypical symptoms.
4. Symptom 2: no symptoms vs abnormal.
5. ECG: normal, nondiagnostic vs abnormal.

**Table 4.** Comparison of ETT Variables by NLP Identified ETT Results

ETT Variables	Normal	Abnormal	Equivocal	Nondiagnostic	P Value*	Total
No. (%)	3908 (75)	310 (5.9)	344 (6.6)	652 (12.5)		5214 (100)
Days between ED and ETT	1 (1, 5)	1 (1, 3)	1 (1, 5.5)	1 (1, 3)	<0.001	5214 (100)
Protocol—standard Bruce	3745 (95.8)	298 (96.1)	326 (94.8)	562 (86.2)	<0.001	4931 (94.6)
Exercise time, min	8.8 (6.6, 10)	7.2 (6, 9.1)	7.6 (6, 9.4)	6.4 (4.3, 8.4)	<0.001	5079 (97.4)
<b>BP</b>						
Resting SBP	128 (117, 141)	131 (118, 142)	132 (120, 144.5)	133 (120, 146)	<0.001	4780 (91.7)
Resting DBP	80 (72, 86)	79 (70, 88)	80 (72, 88)	78 (70, 84)	<0.001	4781 (91.7)
Resting pulse pressure	48 (40, 58)	50 (41, 61)	52 (41, 61)	54 (44, 66)	<0.001	4780 (91.7)
Maximum SBP	178 (160, 196)	180 (162, 199)	181 (162, 198)	174 (155, 196)	0.005	4780 (91.7)
Maximum DBP	80 (70, 88)	79 (70, 87)	80 (71, 88)	80 (69, 87)	0.2	4780 (91.7)
Maximum pulse pressure	98 (80, 117)	100.5 (82, 120.5)	100 (83, 118)	94 (78, 115)	0.03	4780 (91.7)
SBP change	50 (36, 63)	48 (33, 65)	49 (36, 60)	41 (28, 58)	<0.001	4586 (88)
Hypertensive	1342 (34.3)	98 (31.6)	126 (36.6)	199 (30.5)	0.14	1765 (33.9)
Hypertensive (diastolic)	693 (17.7)	49 (15.8)	64 (18.6)	115 (17.6)	<0.001	921 (17.7)
Hypertensive (systolic)	828 (21.2)	65 (21)	86 (25)	123 (18.9)	<0.001	1102 (21.1)
Hypotensive	3 (0.1)	1 (0.3)	1 (0.3)	3 (0.5)	0.04 <sup>†</sup>	8 (0.2)
Low SBP peak	208 (5.3)	23 (7.4)	19 (5.5)	63 (9.7)	0.001	313 (6)
<b>HR</b>						
Resting HR	74 (65, 83)	69 (63, 78)	73 (64, 82)	67 (60, 76)	<0.001	4822 (92.5)
Maximum HR	155 (146, 166)	150 (139, 160)	153 (141, 162)	126 (114, 139)	<0.001	4939 (94.7)
MPHR	94 (89, 100)	90 (86, 98)	92 (87, 98)	78 (72, 83)	<0.001	5170 (99.2)
Chronotropic incompetence	852 (21.8)	108 (34.8)	109 (31.7)	491 (75.3)	<0.001	1560 (29.9)
METs					<0.001	5100 (97.8)
≤7	745 (19.1)	92 (29.7)	101 (29.4)	291 (44.6)		1229 (23.6)
7 to 10	926 (23.7)	76 (24.5)	78 (22.7)	153 (23.5)		1233 (23.6)
>10	2178 (55.7)	135 (43.5)	160 (46.5)	165 (25.3)		2638 (50.6)
Symptom					<0.001	5214 (100)
Abnormal chest pain	113 (2.9)	73 (23.5)	24 (7)	41 (6.3)		251 (4.8)
Atypical angina	264 (6.8)	52 (16.8)	36 (10.5)	85 (13)		437 (8.4)
Atypical symptoms	279 (7.1)	21 (6.8)	29 (8.4)	93 (14.3)		422 (8.1)
No symptoms	3252 (83.2)	164 (52.9)	255 (74.1)	433 (66.4)		4104 (78.7)
ECG finding					<0.001	5199 (99.7)
Abnormal	47 (1.2)	152 (49)	74 (21.5)	35 (5.4)		308 (5.9)
Nondiagnostic	300 (7.7)	28 (9)	105 (30.5)	70 (10.7)		503 (9.6)
Normal	3561 (91.1)	130 (41.9)	165 (48)	532 (81.6)		4388 (84.2)
Reason for stopping <sup>‡</sup>					<0.001	
Target HR achieved	3489 (71.3)	229 (54.3)	298 (66.4)	482 (51.4)		4498 (67.1)
Noncardiac	268 (5.5)	31 (7.3)	34 (7.6)	143 (15.3)		476 (7.1)
Abnormal BP response	108 (2.2)	7 (1.7)	13 (2.9)	39 (4.2)		167 (2.5)
Dyspnea	271 (5.5)	44 (10.4)	31 (6.9)	80 (8.5)		426 (6.4)
Chest pain	163 (3.3)	61 (14.5)	20 (4.5)	55 (5.9)		299 (4.5)
Missing	592 (12.1)	50 (11.8)	53 (11.8)	138 (14.7)		833 (12.4)

Continuous variables are shown as median (25th, 75th percentiles). Data are presented as number (percentage) unless otherwise indicated. BP indicates blood pressure; DBP, diastolic blood pressure; ED, emergency department; ETT, exercise treadmill test; HR, heart rate; METs, metabolic equivalents; MPHR, maximum predicted heart rate; NLP, natural language processing; SBP, systolic blood pressure.

\*Chi-square test was used for categorical variables and Wilcoxon test was used for continuous variables.

<sup>†</sup>Fisher exact test.

<sup>‡</sup>Reason for stopping allows multiple values per report.

**Table 5.** Thirty-Day Major Adverse Cardiac Outcomes Stratified by NLP Identified Treadmill Test Results After an ED Visit for Suspected Acute Coronary Syndrome

30-d Outcomes	NLP Identified ETT Results								P Value*	Total	
	Normal		Abnormal		Equivocal		Nondiagnostic				
	No.	% (95% CI)	No.	% (95% CI)	No.	% (95% CI)	No.	% (95% CI)		No.	% (95% CI)
MACE	3	0.08 (0–0.16)	31	10 (6.66–13.34)	7	2.03 (0.54–3.53)	7	1.07 (0.28–1.86)	<0.001	48	0.92 (0.66–1.18)
Death	0	0 (0–0)	1	0.32 (0–0.95)	0	0 (0–0)	0	0 (0–0)	0.06	1	0.02 (0–0.06)
AMI	3	0.08 (0–0.16)	5	1.61 (0.21–3.02)	2	0.58 (0–1.38)	5	0.77 (0.1–1.44)	<0.001	15	0.29 (0.14–0.43)
CABG	0	0 (0–0)	16	5.16 (2.7–7.62)	1	0.29 (0–0.86)	2	0.31 (0–0.73)	<0.001	19	0.36 (0.2–0.53)
Revascularization	2	0.05 (0–0.12)	12	3.87 (1.72–6.02)	5	1.45 (0.19–2.72)	3	0.46 (0–0.98)	<0.001	22	0.42 (0.25–0.6)
Death or AMI	3	0.08 (0–0.16)	6	1.94 (0.4–3.47)	2	0.58 (0–1.38)	5	0.77 (0.1–1.44)	<0.001	16	0.31 (0.16–0.46)

AMI indicates acute myocardial infarction; CABG, coronary artery bypass grafting; ED, emergency department; ETT, exercise treadmill test; MACE, major adverse cardiac events (which included cardiovascular death, acute myocardial infarction, coronary artery bypass grafting, and coronary revascularization); NLP, natural language processing.

\*Fisher exact test.

algorithms developed in our institution have been successful in other institutions.<sup>27,28</sup>

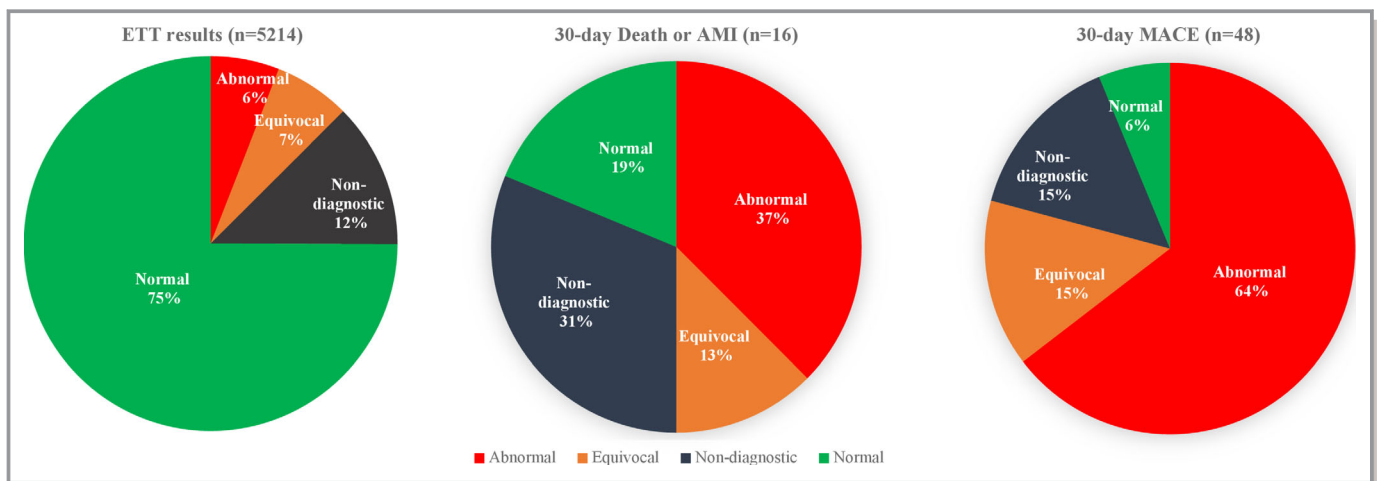
Our results were further validated by the varying association of each ETT result category with 30-day AMI or death. These findings indicate that NLP can be used to facilitate future research and gain better understanding of the benefits and risks of ETT. This may help physicians to identify patients who might benefit from the use of ETT.

Prior studies categorized results into 2 categories (normal and abnormal)<sup>22</sup> or included a third category of “inconclusive,” which combined equivocal and nondiagnostic results.<sup>19,20,23,29</sup> However, our study demonstrated that there are significant differences between “equivocal” and “nondiagnostic” results. Patients with equivocal and nondiagnostic results most closely resembled those with normal and abnormal results, respectively, in baseline characteristics.

Patients with equivocal ETT test results were more likely to have non-normal ECG findings.

Few studies have focused on the prognostic value of ETT in patients with short-term cardiac events referred from the ED with suspected acute coronary syndrome. Compared with a related study composed of a much smaller patient population, our study found lower 30-day death or AMI rates for patients with normal (0.17% versus 0.08%) or ischemic (3.5% versus 1.9%) ETTs but higher rates for those with nondiagnostic (0% versus 0.77%) results.<sup>20</sup> Three-fourths of our study population had normal ETT results, consistent with other reports.<sup>19–21</sup> The overall 30-day death/AMI rate was low (0.31%; 95% CI, 0.16–0.46), which may suggest that patients are sent for stress testing too often and a better pretest risk stratification is needed.

Even within an integrated health system, we identified numerous variations on the format and quality of the ETT



**Figure 2.** Thirty-day MACE stratified by natural language processing–identified treadmill test results after an emergency department visit for suspected acute coronary syndrome. AMI indicates acute myocardial infarction; ETT, exercise treadmill test; MACE, major adverse cardiac events (which included cardiovascular death, acute myocardial infarction, coronary artery bypass grafting, and coronary revascularization).



**Table 6.** Thirty-Day Major Adverse Cardiac Outcomes Stratified by NLP Identified Treadmill Test Results After an ED Visit for Suspected Acute Coronary Syndrome

ETT Results	30-d MACE		30-d Death or AMI	
	No. of Cases	OR (95% CI)*	No. of Cases	OR (95% CI)*
Abnormal vs normal	31:3	125.8 (47.2–466.3)	6:3	23.8 (6.7–100.4)
Equivocal vs normal	7:3	24.8 (7.3–102.5)	2:3	8.1 (1.4–42.0)
Nondiagnostic vs normal	7:3	13.0 (3.8–53.5)	5:3	9.5 (2.5–40.9)

Number of patients in the 4 groups of exercise treadmill test (ETT) results: abnormal=310; equivocal=344; nondiagnostic=652; and normal=3908. AMI indicates acute myocardial infarction; ED, emergency department; MACE, major adverse cardiac events (which included cardiovascular death, acute myocardial infarction, coronary artery bypass grafting, and coronary revascularization); NLP, natural language processing; OR, odds ratio.

\*Logistic regression with Firth penalized maximum likelihood estimation.

reports. While some reports contained the most information in a well-formed format, as shown in the sample ETT report (Data S5), others had missing data elements, section heads, and punctuation. NLP also identified incorrect and missing information in the reports (Table S5). In addition to its usage in research studies, this method can be integrated into the EHR system to improve the quality of ETT reports, thus improving clinical decision support and care coordination for patients undergoing ETT. Proper treatment and follow-up of patients undergoing ETT are essential to reduce the risk of future cardiac events. NLP's ability to extract useful information from unstructured data available in the EHR may enable more efficient, economically feasible, large-scale applications using ETT data among diverse systems.

There were significant differences in the majority of extracted variables between ETT result groups. These variables have been reported to have additional diagnostic or prognostic values in addition to the ETT result.<sup>30</sup> The Duke Treadmill Score is a weighted score combining exercise time, ST change, and exercise-induced angina.<sup>31</sup> It has been used as a risk-stratification tool and to predict 5-year mortality. However, it was developed for ETT under the Bruce protocol and did not include other ETT variables such as METs, HR, or blood pressure. The FIT Treadmill Score was derived by combining age, sex, maximum predicted HR, and METs.<sup>32</sup> It was used to predict 10-year mortality and did not include other variables such as ECG, HR, or blood pressure. There are a lack of population-based studies on short-term outcomes prediction following ETT.<sup>33</sup> A much larger study population is required for short-term outcome prediction because of the low incidence rate. The risk models were also commonly linear equations derived by Cox regression models. In the era of artificial intelligence and big data, better machine learning methods have been available to train on a large volume of data efficiently.<sup>34</sup> The new machine learning methods are also able to deal with the imbalanced data such as the low positive cardiac outcomes following ETTs. The NLP

algorithm developed in this study facilitates the development of a more robust risk score system using statistical and machine learning methods. Such a system may provide better prognostic value than the raw ETT results.

### Study Strengths and Limitations

To the best of our best knowledge, this is the largest study on the association of ETT results with short-term cardiac event rates. We found that most patients are at low risk and have normal ETT results, while those with ischemic, nondiagnostic, or equivocal results have higher risks and warrant future research to help direct clinical management.

Our study population was limited to patients in a large integrated health system presenting to the ED with ETT performed within 30 days. ETTs were also performed for patients in non-ED settings. The automated approach developed in this study does not rely on any specific clinical features unique to our institution. ETT results were mainly based on the treating clinician's interpretations, rather than adjudicated by a core laboratory. However, variations in test interpretation are expected among the clinicians. We limited our analyses to short-term outcomes using only the ETT result since it is often the only information used in clinical decision making.<sup>23</sup> The other variables extracted by the NLP in this study could be used to augment the ETT results for better prediction of short-term outcomes in future studies. Our study focused on the ETT reports, which do not have ECG tracing information. The only structured data we used in the algorithms were the patient's age and sex. Including additional clinical variables will likely enhance short-term outcome prediction. Patients presenting to the ED with ETT have a low rate of short-term cardiac events. Of more than 5000 patients, only 16 had an AMI or died at 30 days (Table 5). In the future, we may reassess these correlations in a larger population.

### Conclusions

We developed and validated an automated NLP algorithm to identify and extract ETT results that performed with high

sensitivity and specificity. We demonstrated that a computational tool could be used to support a population-based study using ETT data otherwise infeasible because of the extensive manual chart review that would be required. The automated identification of ETT variables may facilitate future research to understand the appropriate care strategies for patients who present with suspected acute coronary syndrome in ED settings.

## Acknowledgments

The authors thank the patients of Kaiser Permanente for helping us improve care through the use of information collected through our EHR systems. We also thank Danielle E. Altman, MA, and Stacy J. Park, PhD, for their assistance in organizing and managing this project.

## Sources of Funding

This work was supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health (NIH) under award number R01HL134647. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Dr Natsui was supported by a NIH/National Center for Advancing Translational Sciences UCLA CTSI grant (TL1TR001883). Dr Ferencik was supported by an American Heart Association Fellow-to-Faculty Award (13FTF16450001).

## Disclosures

Dr Sun was a consultant for Medtronic. The remaining authors have no disclosures to report.

## References

- Brooker JA, Hastings JW, Major-Monfried H, Maron CP, Winkel M, Wijeratne HR, Fleischman W, Weingart S, Newman DH. The association between medicolegal and professional concerns and chest pain admission rates. *Acad Emerg Med*. 2015;22:883–886.
- Fihn SD, Gardin JM, Abrams J, Berra K, Blankenship JC, Dallas AP, Douglas PS, Foody JM, Gerber TC, Hinderliter AL, King SB III, Kligfield PD, Krumholz HM, Kwong RY, Lim MJ, Linderbaum JA, Mack MJ, Munger MA, Prager RL, Sabik JF, Shaw LJ, Sikkema JD, Smith CR Jr, Smith SC Jr, Spertus JA, Williams SV; American College of Cardiology F. 2012 ACCF/AHA/ACP/AATS/PCNA/SCAI/STS guideline for the diagnosis and management of patients with stable ischemic heart disease: executive summary: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines, and the American College of Physicians, American Association for Thoracic Surgery, Preventive Cardiovascular Nurses Association, Society for Cardiovascular Angiography and Interventions, and Society of Thoracic Surgeons. *Circulation*. 2012;126:3097–3137.
- Prasad V, Cheung M, Cifu A. Chest pain in the emergency department: the case against our current practice of routine noninvasive testing. *Arch Intern Med*. 2012;172:1506–1509.
- Redberg RF. Stress testing in the emergency department: not which test but whether any test should be done. *JAMA Intern Med*. 2015;175:436.
- Foy AJ, Liu G, Davidson WR Jr, Sciamanna C, Leslie DL. Comparative effectiveness of diagnostic testing strategies in emergency department patients with chest pain: an analysis of downstream testing, interventions, and outcomes. *JAMA Intern Med*. 2015;175:428–436.
- Venkatesh AK, Geisler BP, Gibson Chambers JJ, Baugh CW, Bohan JS, Schuur JD. Use of observation care in US emergency departments, 2001 to 2008. *PLoS One*. 2011;6:e24326.
- Sabbatini AK, Nallamothu BK, Kocher KE. Reducing variation in hospital admissions from the emergency department for low-mortality conditions may produce savings. *Health Aff (Millwood)*. 2014;33:1655–1663.
- Heston TF. Letter by Heston regarding article, “comparative effectiveness of exercise electrocardiography with or without myocardial perfusion single photon emission computed tomography in women with suspected coronary artery disease: results from the what is the optimal method for ischemia evaluation in women (WOMEN) trial”. *Circulation*. 2012;125:e933; author reply e932–935.
- Zheng C, Rashid N, Wu YL, Koblick R, Lin AT, Levy GD, Cheatham TC. Using natural language processing and machine learning to identify gout flares from electronic clinical notes. *Arthritis Care Res (Hoboken)*. 2014;66:1740–1748.
- Maddox TM, Matheny MA. Natural language processing and the promise of big data: small step forward, but many miles to go. *Circ Cardiovasc Qual Outcomes*. 2015;8:463–465.
- Zheng C, Rashid N, Koblick R, An J. Medication extraction from electronic clinical notes in an integrated health system: a study on aspirin use in patients with nonvalvular atrial fibrillation. *Clin Ther*. 2015;37:2048–2058.e2042.
- Xie F, Zheng C, Yuh-Jer Shen A, Chen W. Extracting and analyzing ejection fraction values from electronic echocardiography reports in a large health maintenance organization. *Health Informatics J*. 2017;23:319–328.
- An J, Niu F, Zheng C, Rashid N, Mendes RA, Dills D, Vo L, Singh P, Bruno A, Lang DT, Le PT, Jazdzewski KP, Aranda G Jr. Warfarin management and outcomes in patients with nonvalvular atrial fibrillation within an integrated health care system. *J Manag Care Spec Pharm*. 2017;23:700–712.
- Backus BE, Six AJ, Kelder JC, Mast TP, van den Akker F, Mast EG, Monnick SH, van Tooren RM, Doevendans PA. Chest pain in the emergency room: a multicenter validation of the HEART Score. *Crit Pathw Cardiol*. 2010;9:164–169.
- Sharp AL, Wu YL, Shen E, Redberg R, Lee MS, Ferencik M, Natsui S, Zheng C, Kawatkar A, Gould MK, Sun BC. The HEART score for suspected acute coronary syndrome in U.S. emergency departments. *J Am Coll Cardiol*. 2018;72:1875–1877.
- Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC, Saunders LD, Beck CA, Feasby TE, Ghali WA. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care*. 2005;43:1130–1139.
- Sharp AL, Baecker AS, Shen E, Redberg R, Lee MS, Ferencik M, Natsui S, Zheng C, Kawatkar A, Gould MK, Sun BC. Effect of a HEART care pathway on chest pain management within an integrated health system. *Ann Emerg Med*. 2019;74:171–180.
- Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. *J Biomed Inform*. 2014;48:193–204.
- Christman MP, Bittencourt MS, Hulten E, Saksena E, Hainer J, Skali H, Kwong RY, Forman DE, Dorbala S, O’Gara PT, Di Carli MF, Blankstein R. Yield of downstream tests after exercise treadmill testing: a prospective cohort study. *J Am Coll Cardiol*. 2014;63:1264–1274.
- Amsterdam EA, Kirk JD, Diercks DB, Lewis WR, Turnipseed SD. Immediate exercise testing to evaluate low-risk patients presenting to the emergency department with chest pain. *J Am Coll Cardiol*. 2002;40:251–256.
- Diercks DB, Hollander JE, Sites F, Kirk JD. Derivation and validation of a risk stratification model to identify coronary artery disease in women who present to the emergency department with potential acute coronary syndromes. *Acad Emerg Med*. 2004;11:630–634.
- Fletcher GF, Ades PA, Kligfield P, Arena R, Balady GJ, Bittner VA, Coke LA, Fleg JL, Forman DE, Gerber TC, Gulati M, Madan K, Rhodes J, Thompson PD, Williams MA; American Heart Association Exercise, Cardiac Rehabilitation, and Prevention Committee of the Council on Clinical Cardiology, Council on Nutrition, Physical Activity and Metabolism, Council on Cardiovascular and Stroke Nursing, and Council on Epidemiology and Prevention. Exercise standards for testing and training: a scientific statement from the American Heart Association. *Circulation*. 2013;128:873–934.
- Amsterdam EA, Kirk JD, Bluemke DA, Diercks D, Farkouh ME, Garvey JL, Kontos MC, McCord J, Miller TD, Morise A, Newby LK, Ruberg FL, Scordo KA, Thompson PD. Testing of low-risk patients presenting to the emergency department with chest pain: a scientific statement from the American Heart Association. *Circulation*. 2010;122:1756–1776.
- Firth D. Bias reduction of maximum likelihood estimates. *Biometrika*. 1993;80:27–38.
- Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159–174.
- Vassar M, Holzmann M. The retrospective chart review: important methodological considerations. *J Educ Eval Health Prof*. 2013;10:12.

27. Zheng C, Yu W, Xie F, Chen W, Mercado C, Sy LS, Qian L, Glenn S, Lee G, Tseng HF, Duffy J, Jackson LA, Daley MF, Crane B, McLean HQ, Jacobsen SJ. The use of natural language processing to identify Tdap-related local reactions at five health care systems in the Vaccine Safety Datalink. *Int J Med Inform.* 2019;127:27–34.
28. Danforth KN, Early MI, Ngan S, Kosco AE, Zheng C, Gould MK. Automated identification of patients with pulmonary nodules in an integrated health system using administrative health plan data, radiology reports, and natural language processing. *J Thorac Oncol.* 2012;7:1257–1262.
29. Barraclough K, Gale CP, Hall R. Assessment of chest pain in a low risk patient: is the exercise tolerance test obsolete? *BMJ.* 2015;350:h1905.
30. Ashley EA, Myers J, Froelicher V. Exercise testing in clinical medicine. *Lancet.* 2000;356:1592–1597.
31. Mark DB, Shaw L, Harrell FE Jr, Hlatky MA, Lee KL, Bengtson JR, McCants CB, Califf RM, Pryor DB. Prognostic value of a treadmill exercise score in outpatients with suspected coronary artery disease. *N Engl J Med.* 1991;325:849–853.
32. Ahmed HM, Al-Mallah MH, McEvoy JW, Nasir K, Blumenthal RS, Jones SR, Brawner CA, Keteyian SJ, Blaha MJ. Maximal exercise testing variables and 10-year survival: fitness risk score derivation from the FIT Project. *Mayo Clin Proc.* 2015;90:346–355.
33. Jonas DE, Reddy S, Middleton JC, Barclay C, Green J, Baker C, Asher GN. Screening for cardiovascular disease risk with resting or exercise electrocardiography: evidence report and systematic review for the US Preventive Services Task Force. *JAMA.* 2018;319:2315–2328.
34. Murphy KP. *Machine Learning: A Probabilistic Perspective.* Cambridge, MA: The MIT Press; 2012.

# **Supplemental Material**

## **Data S1.**

### **International Classification of Diseases (ICD) codes and Current Procedural Terminology (CPT) codes used in creating the study population**

**Definition of noninvasive cardiac testing:** 1) A completed ED referral to a KPSC cardiology department with CPT code 200267; or 2) an outpatient visit with CPT code for a stress ECG test (CPT: 93015, 93018).

Due to KPSC adoption of ICD-10 starting October 2015, the use of ICD-9 codes covers the period from June through September 2015, and ICD-10 codes cover the remaining period.

**Definition of AMI:** ICD-9 4100, 4101, 4102, 410.10, 410.11, 410.12, 410.20, 410.21, 410.22, 410.30, 410.31, 410.32, 410.40, 410.41, 410.42, 410.50, 410.51, 410.52, 410.60, 410.61, 410.62, 410.70, 410.71, 410.72, 410.80, 410.81, 410.82, 410.90, 410.91, 410.92 and ICD-10 I21, I22 codes.

**Revascularization and CABG** were defined using extensive lists of ICD-9, ICD-10, and CPT codes, which we do not present here for brevity but can provide upon request.

## Data S2.

### NLP extracted and derived variables from ETT reports

ETT reports include a rich set of diagnostic and prognosis information.<sup>1</sup> We extracted the following variables from ETT reports:

*Study protocol* defines the speed and inclination at specified time intervals. Different protocols were used based on patients' physical conditions.<sup>1,2</sup> Study protocols we extracted and grouped into two categories:

- **Standard Bruce:** Bruce
- **Other protocols:** modified Bruce, Astrand, Balke, Cornell, Ellestad, Naughton, manual, etc.

*Exercise time* is denoted in seconds which measures functional capacity and is one of the most influential prognostic factors.<sup>3,4</sup>

*Blood pressure (BP)* is measured before and during the ETT. The blood pressure response to exercise can be used to prognosticate the risk of cardiovascular disease.<sup>2,3</sup>

*Heart rate (HR)* variables include resting HR, maximum HR, achieved MPHR (maximal predicted heart rate).<sup>3,5</sup>

*Metabolic equivalents (METS)* is a functional capacity measurement calculated based on the speed and grade of the ETT.<sup>6</sup> METS was shown to be a powerful predictor of mortality.<sup>7</sup> The METS extracted is a numerical variable.

**ECG ST change** is the ST-segment depression information extracted from ETT reports.

*Symptoms during exercise* and *reasons for stopping exercise* refer to patient-reported symptoms and may have prognostic value for future cardiac events or death.<sup>2,8,9</sup>

**Symptom** was grouped into three categories:

- Abnormal: such as angina, typical angina, chest pressure

- Atypical angina: such as atypical angina, burning chest pain, arrhythmia, atrioventricular block, bradycardia, left /right bundle branch block, tachycardia, ST, and blood pressure abnormality
- Atypical symptoms: such as fatigue, shortness of breath, dizziness, hypertensive response, other atypical symptoms

**Reason for stopping** was grouped into five categories:

- Endpoint: such as fatigue, reach study endpoint
- Noncardiac: such as exercise intolerance, safety, headache, dizziness, nausea
- BP response: such as abnormal BP responses
- Dyspnea: such as shortness of breath, cough
- Cardiac: such as chest pain, arrhythmia, atrioventricular block, cardiac arrhythmia, left bundle branch block, paroxysmal atrial tachycardia, right bundle branch block, sinus bradycardia, sinus tachycardia, ventricular bigeminy, and other cardiac reasons

***Clinician assessment*** is the overall impression of the ETT stated by the clinician, which was classified into four categories: normal, non-diagnostic, equivocal, and abnormal.

From the extracted variables, we further derived the following variables from ETT reports:

***ETT result*** is the final result of the ETT synthesized by the NLP algorithm based on clinical assessment and other extracted ETT variables. ETT result was classified into four categories: normal, non-diagnostic, equivocal, and abnormal. The last three categories were also referred to as non-normal in this study.

***ECG result*** refers to ECG changes suggestive of ischemia<sup>2</sup>. Previous studies showed ischemic ECG to be a strong predictor of cardiac events.<sup>10</sup> The ST-segment depression

information extracted from ETT reports was used to derive the final ECG results by combining with the extracted ECG assessment (Supplemental Method 2). Results were defined as normal, non-diagnostic, and abnormal.

***Chronotropic index*** is a heart rate related variable defined as  $(\text{maximum HR} - \text{resting HR}) / ((220 - \text{age}) - \text{resting HR})$

Chronotropic incompetence is defined as Chronotropic index  $< 0.8$

***Blood pressure related variables:***

Additional variables such as hypertensive and hypotensive response, low SBP (systolic blood pressure) peak were derived based on the definition described below:

- SBP: systolic blood pressure
- DBP: diastolic blood pressure
- Resting pulse pressure: resting SBP – resting DBP
- Maximum pulse pressure: maximum SBP – maximum DBP
- SBP change: maximum SBP - resting SBP
- Hypertensive (diastolic) response:  $(\text{maximum DBP} - \text{resting DBP}) > 10 \text{ mm Hg}$  or  $\text{maximum DBP} \geq 100 \text{ mm Hg}$
- Hypertensive (systolic) response: maximum SBP  $\geq 210 \text{ mm Hg}$  for men and  $\geq 190 \text{ mm Hg}$  for women
- Hypertensive response: hypertensive (diastolic) response or hypertensive (diastolic) response
- Hypotensive response: maximum SBP  $< \text{resting SBP}$
- Low systolic peak: maximum SBP  $< 140 \text{ mm Hg}$  or  $(\text{maximum SBP} - \text{resting SBP}) < 10$



## Data S3.

### Algorithm to derive the result of the treadmill test report

#### Inputs from NLP extracted variables:

- **Assessment:** Normalized ETT result based on clinician's assessment (normal, abnormal, equivocal, non-diagnostic)
- **MHR:** Maximum heart rate achieved in the test
- **MPHR:** Maximum predicted heart rate achieved in the test (%)
- **ECG:** Final ECG component derived by NLP (normal, abnormal, equivocal)

#### Output:

- **ETT\_final:** Final ETT test result derived by NLP (normal, abnormal, equivocal, non-diagnostic, etc.)

if MPHR is not found:

    if age and MHR are available:

$$\text{MPHR} = \text{MHR} * 100 / (220 - \text{age})$$

if Assessment is found:

**ETT\_final** = Assessment

else if MPHR is found:

    if (MPHR  $\geq$  85): **ETT\_final** = ECG

    else if (MPHR < 85) and (ECG is not normal): **ETT\_final** = ECG

    else if (MPHR < 85): **ETT\_final** = non-diagnostic

## **Data S4.**

### **Sample list of descriptors used to identify subjective assessment in the ETT reports**

#### **Abnormal**

abnormal, abn, high risk, positive for, complained, c/o

#### **Equivocal**

equivocal, borderline, cannot be ruled out, concerning for, could be considered, intermediate risk, non-specific, possible, seems to be, suggestive, remain a consideration

#### **Non-diagnostic**

non-diagnostic, did not achieve, failure to achieve, inconclusive, not decisive, non dx, not diagnostic for, submaximal, not performed, aborted, cannot perform, not done, unable to walk, cancelled, deferred, postponed

#### **Normal**

normal, does not meet ischemia criteria, lack of, least likely, low risk, low suspicion, negative, no evidence of, no stress-induced, non ischemic, unlikely, unremarkable

**Data S5.**

**Sample treadmill test report**

**TREADMILL EXERCISE STRESS TEST (BRUCE PROTOCOL)**

Reason for Test: Chest tightness and felt SOB with numbness both arms and legs.

Resting EKG:SR at 60, 1st degree AVHB.

Target HR:85 %

Max Predicted Heart Rate:123 bpm

Pre-test symptoms

**MIN MPH % GRADE HR BP COMMENTS**

0 Resting 60 80/52 supine Left arm denied chest sx/denied dizziness; baseline mild head discomfort.

Baseline symptoms

1 1.7 10 81 Denied chest sx/denied dizziness; baseline mild head discomfort.

2 1.7 10 88 " "

3 1.7 10 90 87/53 " " PACs

4 2.5 12 96 " "

5 2.5 12 102 tired/pt requested to slow down; chest pressure

7-8/10; denied dizziness; increased baseline head discomfort.

During exercise symptoms

**MIN POST EXERCISE HR BP COMMENTS**

Recovery symptoms

1 91 84/41 chest pressure 5-6/10; denied dizziness; baseline mild head discomfort.

3 77 110/55 chest pressure resolving; denied dizziness; baseline mild head discomfort.

6 74 100/54 chest pressure resolving; denied dizziness; baseline mild head discomfort.

9 71 87/52 chest pressure resolved; denied dizziness; baseline mild head discomfort.

**INTERPRETATION:**

Test Stopped after 6 min. 48 sec. of exercise.

pt requested to slow down, feel tired.



**Neighboring sentences**

Achieved: 78 % PMHR (9.4 METS)

Peak HR achieved: 114 bpm

Peak BP achieved: 110 / 55 mmHg

EX capacity: below average

B/P response: baseline low blood pressure w/o SX noted.

CP: chest pressure; denied dizziness; increased baseline head discomfort at peak exercise.

ST changes:

Pre ex - none

During - 0.5 mm horizontal-upslope ST depression at peak exercise

Post ex - return to baseline



**Section**

Ectopy / Arrhythmia(s):

PAC's

**IMPRESSION:**

Equivocal TMST study with complaint of chest discomfort at peak exercise but no ischemic ECG change at less than 85% MPMR; better than average cardiovascular performance for age and gender by achieving more than 9 METs denied dizziness; increased baseline head discomfort at peak exercise; baseline low blood pressure w/o SX noted.

---

NLP-extracted information were highlighted in colors.

## Data S6.

### Algorithm to impute study protocol variable of the treadmill test report

```
if (3 >= extime > 0) and (4 >= mets > 0):  
    protocol = 'standard Bruce'  
else if (6 >= extime > 3) and (7 >= mets > 4):  
    protocol = 'standard Bruce'  
else if (9 >= extime > 6) and (10 >= mets > 7):  
    protocol = 'standard Bruce'  
else if (12 >= extime > 9) and (13 >= mets > 10):  
    protocol = 'standard Bruce'  
else if (15 >= extime > 12) and (15 >= mets > 13):  
    protocol = 'standard Bruce'  
else if (18 >= extime > 15) and (18 >= mets > 15):  
    protocol = 'standard Bruce'  
else if (21 >= extime > 18) and (21 >= mets > 18):  
    protocol = 'standard Bruce'
```

We only focused on identifying whether the missing protocol is a standard Bruce protocol.

extime: exercise time in minutes

mets: Metabolic equivalents

## Data S7.

### Algorithm to derive the result of the ECG component of the treadmill test report

#### Inputs from NLP extracted variables:

- **st\_mm**: ST change magnitude in mm
- **st\_direction**: ST change direction
- **st\_text**: ST change text description
- **ECG\_text**: ECG assessment text description

#### Output:

- **ST**: ST change categorical value (normal, abnormal, equivocal)
- **ECG**: Final ECG component derived by NLP (normal, abnormal, equivocal)

#### Step 1: Convert st\_change to st\_cat:

if st\_mm is found:

if (st\_mm  $\geq$  2 mm): **ST** = abnormal

else if (st\_mm  $\geq$  1 mm) and (st\_direction equal 'downsloping'): **ST** = abnormal

else if (st\_mm  $\geq$  1 mm) and (st\_direction equal 'horizontal'): **ST** = abnormal

else if (st\_mm  $\geq$  1 mm) and (st\_direction equal 'upsloping'): **ST** = equivocal

else: **ST** = normal

else if st\_text is found: **ST** = st\_text

#### Step 2: Combine ST with ecg\_text to derive the final ECG:

if both **ECG\_text** and **ST** were found:

set **ECG** as one of the more severe results of **ECG\_text** and **ST**;

else: set **ECG** to the one which was found

**Table S1. ETT variables extracted by NLP.**

<b>NLP extracted variables</b>	<b>Value</b>	<b>In the reference standard</b>
Clinician assessment	Equivocal	Yes
Study protocol	Bruce	Yes
Exercise time (sec)	408	Yes
Reason for stopping	fatigue	Yes
Symptom	abnormal	Yes
METS	9.4	Yes
Maximum BP	110/55	Yes
MPHR	78	Yes
Maximum HR	114	Yes
Resting BP	80/52	No*
Resting HR	60	No*
ECG ST change	0.5 mm horizontal	No*

\*NLP's accuracy was not formally evaluated on these variables because these variables were not manually extracted in the reference standard. However, we manually verified the NLP results on these variables and confirmed that their accuracies are similar to other variables.



**Table S2. ETT variables derived based on NLP extracted information.**

<b>NLP derived variables</b>	<b>Value</b>	<b>In the reference standard</b>
ETT result	Non-diagnostic, Equivocal	Yes
ECG result	Non-ischemic	Yes
Chronotropic index	0.64	No*
Resting SBP	80	No*
Resting DBP	52	No*
Resting pulse pressure	28	No*
Maximum SBP	110	No*
Maximum DBP	55	No*
Maximum pulse pressure	55	No*
Hypertensive response	No	No*
SBP change	30	No*
Hypertensive response	No	No*
Hypertensive (diastolic) response	No	No*
Hypertensive (systolic) response	No	No*
Hypotensive response	No	No*
Low systolic pressure peak	Yes	No*

\*Some variables were not manually extracted in the reference standard and not formally validated. The majority of them were derived based on the variables that were formally validated (Table 2).

**Table S3. Kappa scores between the two physicians on the validation dataset measured by the treadmill test variables.**

ETT variables	Kappa (95% CI)	Kappa (95% CI)	Kappa (95% CI)
	Reviewer 1 vs. Reviewer 2	Reviewer 1 vs. Reference Standard	Reviewer 2 vs. Reference Standard
Study protocol*	0.66 (0.50-0.82)	0.66 (0.50-0.82)	1.0 (1.0-1.0)
Exercise time	0.79 (0.62-0.97)	0.79 (0.62-0.97)	1.0 (1.0-1.0)
Reasons for stopping*	0.73 (0.58-0.87)	0.77 (0.63- 0.90)	0.95 (0.88-1.0)
Symptom*	0.74 (0.58-0.91)	0.89 (0.78-1.0)	0.83 (0.69-0.97)
Symptom2*	1.0 (1.0-1.0)	1.0 (1.0-1.0)	1.0 (1.0-1.0)
ECG*	0.78 (0.59-0.97)	0.86 (0.70-1.0)	0.90 (0.76-1.0)
METS	0.95 (0.86-1.0)	0.95 (0.86-1.0)	1.0 (1.0-1.0)
Maximum BP	0.88 (0.77-1.0)	0.97 (0.91-1.0)	0.91 (0.81-1.0)
MPHR	0.95 (0.86-1.0)	0.95 (0.86-1.0)	1.0 (1.0-1.0)
Maximum HR	0.92 (0.83-1.0)	0.97 (0.92-1.0)	0.94 (0.87-1.0)
ETT results	0.86 (0.78-0.95)	0.90 (0.82-0.98)	0.95 (0.89-1.0)

\*For evaluation purposes, the results of these multicategory variables were dichotomized into two categories:

- Study protocol: standard Bruce protocol vs. other types of study protocols
- Reasons for stopping: target heart rate achieved vs. other reasons
- Symptom: no symptoms vs. (abnormal, atypical angina, atypical symptoms)
- Symptom2: no symptoms vs. abnormal
- ECG: (normal, non-diagnostic) vs. abnormal

ETT = exercise treadmill test; BP = blood pressure; ECG = electrocardiogram; HR = heart rate;

METS = metabolic equivalents; MPHR = maximum predicted heart rate; NLP = natural

language processing; NPV = negative predictive value; PPV = positive predictive value.

**Table S4. Troponin values by ETT results.**

<b>Troponin values (ng/ml)</b>	<b>Normal</b>	<b>Abnormal</b>	<b>Equivocal</b>	<b>Non- diagnostic</b>	<b><i>p</i> Value</b>	<b>Total</b>
N (%)	3908 (75)	310 (5.9)	344 (6.6)	652 (12.5)	< 0.0001	5214 (100)
< 0.02	3635 (93)	268 (86.5)	317 (92.2)	568 (87.1)		4788 (91.8)
0.02 – 0.5	271 (6.9)	42 (13.5)	27 (7.8)	83 (12.7)		423 (8.1)
> 0.5	2 (0.1)	0 (0)	0 (0)	1 (0.2)		3 (0.1)

**Table S5. Number of conflicted or missing cases for selected variables.**

<b>ETT variables</b>	<b>Missed cases</b>
	<b>N (%)</b>
Study protocol	89 (1.7%)
Exercise time*	135 (2.6%)
Resting BP	434 (8.3%)
Maximum BP	434 (8.3%)
Resting HR†	392 (7.5%)
Maximum HR‡	275 (5.3)
MPHR‡	44 (0.8%)
METS§	114 (2.2%)
ECG fining	15 (0.3%)
Reason for stopping	833 (12.4%)
Clinician assessment	126 (2.4%)

\* Included cases where the difference between conflicted exercise time is more than 1 minute

† Included cases where the difference between HR is more than 10

‡ Included cases where the difference between MPHHR is more than 5

§ Included cases where the difference between METS is more than 1

Maximum HR has more substantial numbers of conflicted cases since HR is often documented multiple times in the ETT reports. It does The NLP algorithm chose the largest value as the final Maximum HR for these conflicted cases.

## SUPPLEMENTAL REFERENCES:

1. Ashley EA, Myers J, Froelicher V. Exercise testing in clinical medicine. *Lancet*. 2000;356:1592-1597.
2. Fletcher GF, Ades PA, Kligfield P, Arena R, Balady GJ, Bittner VA, Coke LA, Fleg JL, Forman DE, Gerber TC, Gulati M, Madan K, Rhodes J, Thompson PD, Williams MA, American Heart Association Exercise CR, Prevention Committee of the Council on Clinical Cardiology CoNPA, Metabolism CoC, Stroke N, Council on E, Prevention. Exercise standards for testing and training: A scientific statement from the american heart association. *Circulation*. 2013;128:873-934.
3. Miller TD. Exercise treadmill test: Estimating cardiovascular prognosis. *Cleve. Clin. J. Med*. 2008;75:424-430.
4. Blair SN, Kohl HW, 3rd, Barlow CE, Paffenbarger RS, Jr., Gibbons LW, Macera CA. Changes in physical fitness and all-cause mortality. A prospective study of healthy and unhealthy men. *JAMA*. 1995;273:1093-1098.
5. Astrand I. Aerobic work capacity in men and women with special reference to age. *Acta Physiol. Scand. Suppl*. 1960;49:1-92.
6. Jette M, Sidney K, Blumchen G. Metabolic equivalents (mets) in exercise testing, exercise prescription, and evaluation of functional capacity. *Clin. Cardiol*. 1990;13:555-565.
7. Myers J, Prakash M, Froelicher V, Do D, Partington S, Atwood JE. Exercise capacity and mortality among men referred for exercise testing. *N. Engl. J. Med*. 2002;346:793-801.

8. Abidov A, Rozanski A, Hachamovitch R, Hayes SW, Aboul-Enein F, Cohen I, Friedman JD, Germano G, Berman DS. Prognostic significance of dyspnea in patients referred for cardiac stress testing. *N. Engl. J. Med.* 2005;353:1889-1898.
9. Amsterdam EA, Kirk JD, Bluemke DA, Diercks D, Farkouh ME, Garvey JL, Kontos MC, McCord J, Miller TD, Morise A, Newby LK, Ruberg FL, Scordo KA, Thompson PD. Testing of low-risk patients presenting to the emergency department with chest pain: A scientific statement from the american heart association. *Circulation.* 2010;122:1756-1776.
10. Bourque JM, Beller GA. Value of exercise eeg for risk stratification in suspected or known cad in the era of advanced imaging technologies. *JACC Cardiovasc. Imaging.* 2015;8:1309-1321.