ESC European Society of Cardiology
European Heart Journal - Digital Health (2022) **3**, 626–637
https://doi.org/10.1093/ehjdh/ztac047

ORIGINAL ARTICLE

Automated interpretation of stress echocardiography reports using natural language processing

Chengyi Zheng [1,*], Benjamin C. Sun[2], Yi-Lin Wu[1], Maros Ferencik[3], Ming-Sum Lee[4], Rita F. Redberg[5], Aniket A. Kawatkar[1], Visanee V. Musigdilok[1], and Adam L. Sharp[1,6]

[1]Research and Evaluation Department, Kaiser Permanente Southern California, 100 S Los Robles Ave, 2nd Floor, Pasadena, CA 91101, USA; [2]Department of Emergency Medicine and Leonard Davis Institute, University of Pennsylvania, Philadelphia, PA 19104, USA; [3]Oregon Health and Science University, Knight Cardiovascular Institute, Portland, OR 97239, USA; [4]Division of Cardiology, Kaiser Permanente Southern California, Los Angeles Medical Center, Los Angeles, CA 90027, USA; [5]Division of Cardiology, University of California, San Francisco, CA 94143, USA; and [6]Clinical Science Department, Kaiser Permanente Bernard J. Tyson School of Medicine, Pasadena, CA 91101, USA

Received 31 May 2022; revised 8 August 2022; online publish-ahead-of-print 5 September 2022

Aims

Stress echocardiography (SE) findings and interpretations are commonly documented in free-text reports. Reusing SE results requires laborious manual reviews. This study aimed to develop and validate an automated method for abstracting SE reports in a large cohort.

Methods and results

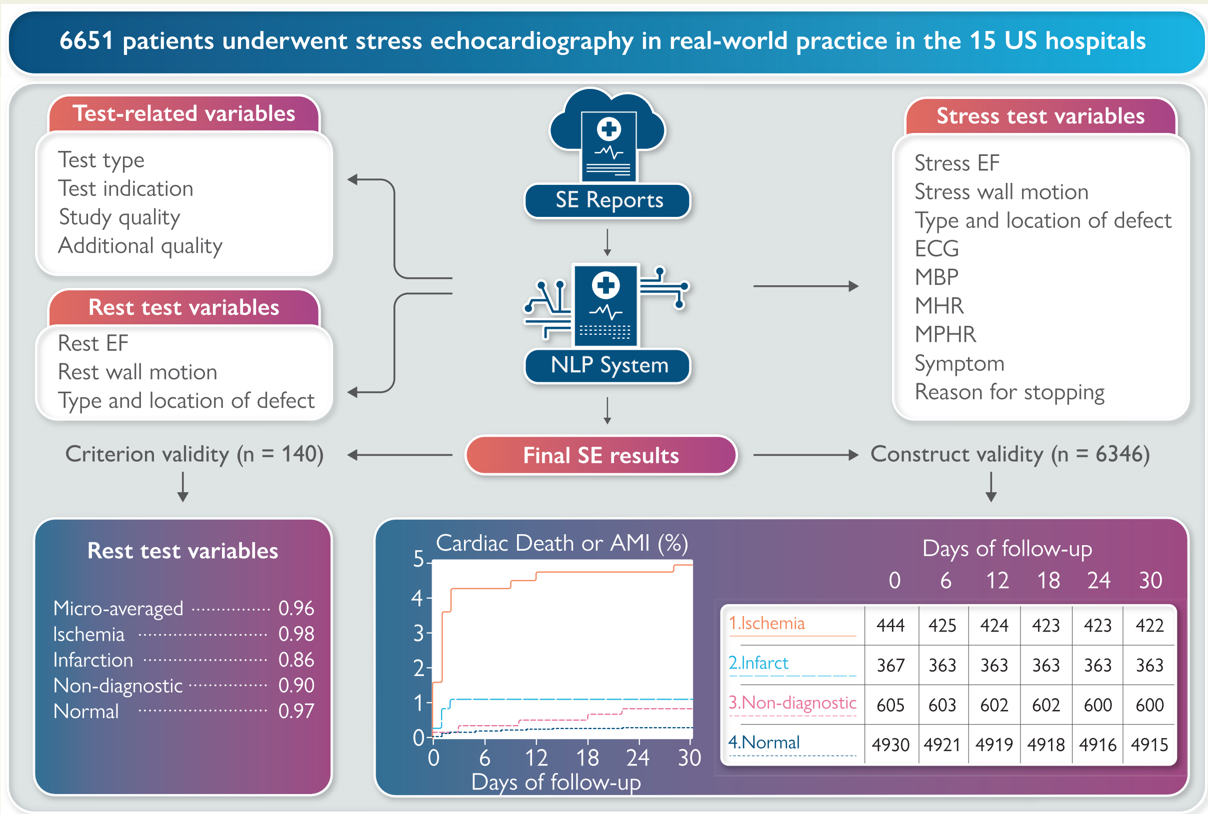
This study included adult patients who had SE within 30 days of their emergency department visit for suspected acute coronary syndrome in a large integrated healthcare system. An automated natural language processing (NLP) algorithm was developed to abstract SE reports and classify overall SE results into normal, non-diagnostic, infarction, and ischaemia categories. Randomly selected reports ($n = 140$) were double-blindly reviewed by cardiologists to perform criterion validity of the NLP algorithm. Construct validity was tested on the entire cohort using abstracted SE data and additional clinical variables. The NLP algorithm abstracted 6346 consecutive SE reports. Cardiologists had good agreements on the overall SE results on the 140 reports: Kappa (0.83) and intraclass correlation coefficient (0.89). The NLP algorithm achieved 98.6% specificity and negative predictive value, 95.7% sensitivity, positive predictive value, and *F*-score on the overall SE results and near-perfect scores on ischaemia findings. The 30-day acute myocardial infarction or death outcomes were highest among patients with ischaemia (5.0%), followed by infarction (1.4%), non-diagnostic (0.8%), and normal (0.3%) results. We found substantial variations in the format and quality of SE reports, even within the same institution.

Conclusions

Natural language processing is an accurate and efficient method for abstracting unstructured SE reports. This approach creates new opportunities for research, public health measures, and care improvement.

* Corresponding author. Tel: 1-626-376-7029, Fax: 626-564-3694, Email: Chengyi.X.Zheng@kp.org

Graphical Abstract



Keywords Natural language processing • Stress echocardiography • Artificial intelligence • Emergency department • Noninvasive stress test • Acute coronary syndrome

Introduction

Acute coronary syndrome (ACS) is the leading cause of death, accounting for >9 million deaths globally in 2019.[1] Stress echocardiography (SE) is a common noninvasive cardiac test to evaluate emergency department (ED) patients with suspected ACS.[2] Stress echocardiography is recommended for patients at intermediate risk of ACS because of its accuracy, moderate cost, and safety. However, its accuracy varies depending on the expertise of the interpreting clinician. Furthermore, there is still considerable disagreement about its cost-effectiveness and appropriate population.[3] Most clinical recommendations and guidelines have relied on meta-analyses of small-scale studies, or studies conducted more than a decade ago.[2,4–8] Large-scale observational studies on SE are becoming possible because clinician-interpreted SE reports become more commonly available. However, manual abstracting of vast amounts of free-text formatted SE reports is time-consuming. Despite the American Society of Echocardiography's recommendations for a standardized echocardiography report two decades ago, most SE reports are still reported in non-standardized formats.[9]

As artificial intelligence becomes more widely adopted in healthcare, natural language processing (NLP) has been used to extract information from unstructured data for many clinical specialties, including cardiology.[10–12] In cardiovascular imaging, NLP has mostly been utilized to extract ejection fractions (EFs) from echocardiography reports.[13] To the best of our knowledge, NLP has never been used to evaluate SE results. Understanding the association between pre-test risk factors and the likelihood of an ischaemic stress test, for example, could help improve the efficiency, affordability, and effectiveness of these diagnostic tests.

This study aims to develop and validate an NLP algorithm to reliably abstract SE reports compared with the cardiologists-created gold standard, and accurately identify SE results in a real-world patient population.

Methods

Study setting

This was a retrospective cohort study among members of Kaiser Permanente Southern California (KPSC), an integrated healthcare organization with over 8500 physicians, 15 hospitals, 234 medical offices, and ∼1 million annual ED visits. Kaiser Permanente Southern California provides prepaid health care to over 4.7 million racially and socio-economically diverse members in KPSC-owned facilities and contracting facilities. All KPSC ED sites used the same troponin lab assay during the study period (Beckman Coulter Access AccuTnI+3). Emergency department physicians at KPSC can order noninvasive cardiac testing as part of the discharge and follow-up plan for patients with suspected acute coronary syndrome (ACS). The KPSC Institutional Review Board approved this study.

Study population

We included all KPSC members aged 18 years or older with an ED visit for suspected ACS triggering a troponin lab order between 01/01/2015 and 12/31/2019, who underwent an SE within 30 days of their visit. We excluded patients who were transferred from a non-KPSC hospital or died in the ED. We also excluded patients without KPSC health plan membership because our data do not accurately capture comorbidities and patient outcomes for non-members. Stress echocardiography studies were identified using Current Procedural Terminology (CPT®) codes (93 350–93 351) with order linked to the index ED visit. Both exercise SE (ESE) and dobutamine SE (DSE) examinations were included in the study. *Figure 1* depicts the flow of participants through the study.

We obtained demographic information including age, sex, and race from administrative records; smoking and family history of coronary artery disease (CAD) from self-reported fields in electronic medical records (EMR); medications from pharmacy records. Body mass index was measured from ED intake documentation or the most recently available visit. Troponin values were extracted from the lab data. HEART (history, electrocardiogram, age, risk factors, troponin) risk stratification scores calculated at the time of the index ED visit were retrieved from the EMR.[14] Comorbidities were defined using the International Classification of Diseases Ninth/Tenth Revision, Clinical Modification (ICD-9/10-CM) codes included in the Elixhauser score.

Stress echocardiography reports

Kaiser Permanente Southern California does not have structured reporting for SE exams. The SE reports were dictated or written by the interpreting physicians in unstructured or free-text formats and saved to the Epic Clarity system running on Oracle Exadata.

Training and validation datasets

By test year, we divided the SE reports for training (2015 and 2018) and validation (2016 and 2017) sampling (*Figure 1*). We chose half of the cases with a major adverse cardiac event (MACE) in the 6 months following SE tests to increase the abnormal cases in the sample data. We created training ($n = 120$) and validation ($n = 150$) datasets using random stratified sampling. Three board-certified cardiologists from three institutions (M.F., M.S.L., and R.F.R.) independently abstracted the SE reports in the training and validation datasets. The cardiologists were blinded to each other's reviews and abstracted solely based on the reports. The results of the cardiologists' review were compared, and conflicts were resolved through discussion and consensus. For the validation data, each cardiologist reviewed 100 cases, and each report was reviewed by two cardiologists. The adjudicated results served as the reference standard against which NLP was evaluated. The reference standard only contained characteristics that were common to both DSE and ESE to maintain a sufficient sample size. We calculated the weighted Cohen's Kappa[15] and the intraclass correlation coefficient (ICC)[16] based on the agreement.

Natural language processing algorithm

We developed an NLP-based algorithm to extract information from the SE reports (*Figure 2*). The core NLP processes have already been discussed in our prior publications on abstracting exercise treadmill test[17] and myocardial perfusion imaging (MPI) reports.[18] Additional technique details for abstracting SE reports are provided in the Supplementary material online. Our main goal was to accurately identify the overall SE test results by categorizing each test into four categories[19,20]:

*Ischaemia**: Stress-induced wall motion abnormality, suggestive of stress-induced ischaemia. *For ischaemia patients, additional infarction finding was also abstracted.

Infarction: No definitive ischaemic finding, but resting wall motion abnormalities and/or significant wall thinning without improvement during stress.

Non-diagnostic: Ischaemia or infarction cannot be ruled out due to the presence of artefacts or sub-optimal test quality, or failure to achieve 85% of the maximum predicted heart rate (MPHR).

Normal: Test quality was sufficient to rule out ischaemia or infarction, and the patient completed the test with an appropriate heart rate.

Besides abstracting the conclusive interpretation, we also captured and synthesized the test results based on the regional wall motion status according to the criteria described in Supplementary material online, *Table S1*.

For cases with ischaemia, we further identified the location and extent of ischaemia (Supplementary material online, *Method S1*). In cases where the extent of ischaemia was not stated in the report, we estimated it based on the number of segments involved. We used the 17-segment model to define the ischaemic extent as small (involving 1–2 segments), medium (3–4 segments), and large (≥5 segments).[19,21] The rest and stress EF results as documented in numerical or ordinal values were separately extracted. The numerical EF values were converted into ordinal values as severely abnormal (<30%), moderately abnormal (≥30 to 39%), mildly abnormal (40–49%), normal (≥50%), and hyperdynamic (>75%). The NLP algorithm extracted other commonly documented information from the SE reports, including stress protocol used, exercise time, maximum heart rate achieved, blood pressure response, maximum rate pressure product (MRPP), exercise capacity, and whether adequate stress was achieved (≥85% MPHR for both exercise and dobutamine, MRPP > 25 000 for exercise stress). Besides the echocardiographic data, clinical and electrocardiographic data were also extracted, including test indication, symptoms reported during stress, reasons for test termination, and ST-segment changes on the stress electrocardiogram.

Criterion validity of natural language processing algorithm

We evaluated the performance of NLP against the reference standard created by double-blinded reviews and consensus among cardiologist reviewers. For multi-class classification with imbalanced data, the micro-averaged scores are the preferred overall performance measures.[22] To compute the micro-averaged scores, we first dichotomized the SE result for each category and counted the numbers of true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Then we calculated the sensitivity, specificity, positive/negative predictive value (PPV/NPV), and *F*-score (i.e. the harmonic mean of sensitivity and PPV) for each dichotomized class. We calculate micro-averaged scores based on the sum of counts of TP, TN, FP, and FN across the classes. Before calculating its performance, the EF value was converted into categorical values of normal (EF ≥ 50) and abnormal (EF < 50).

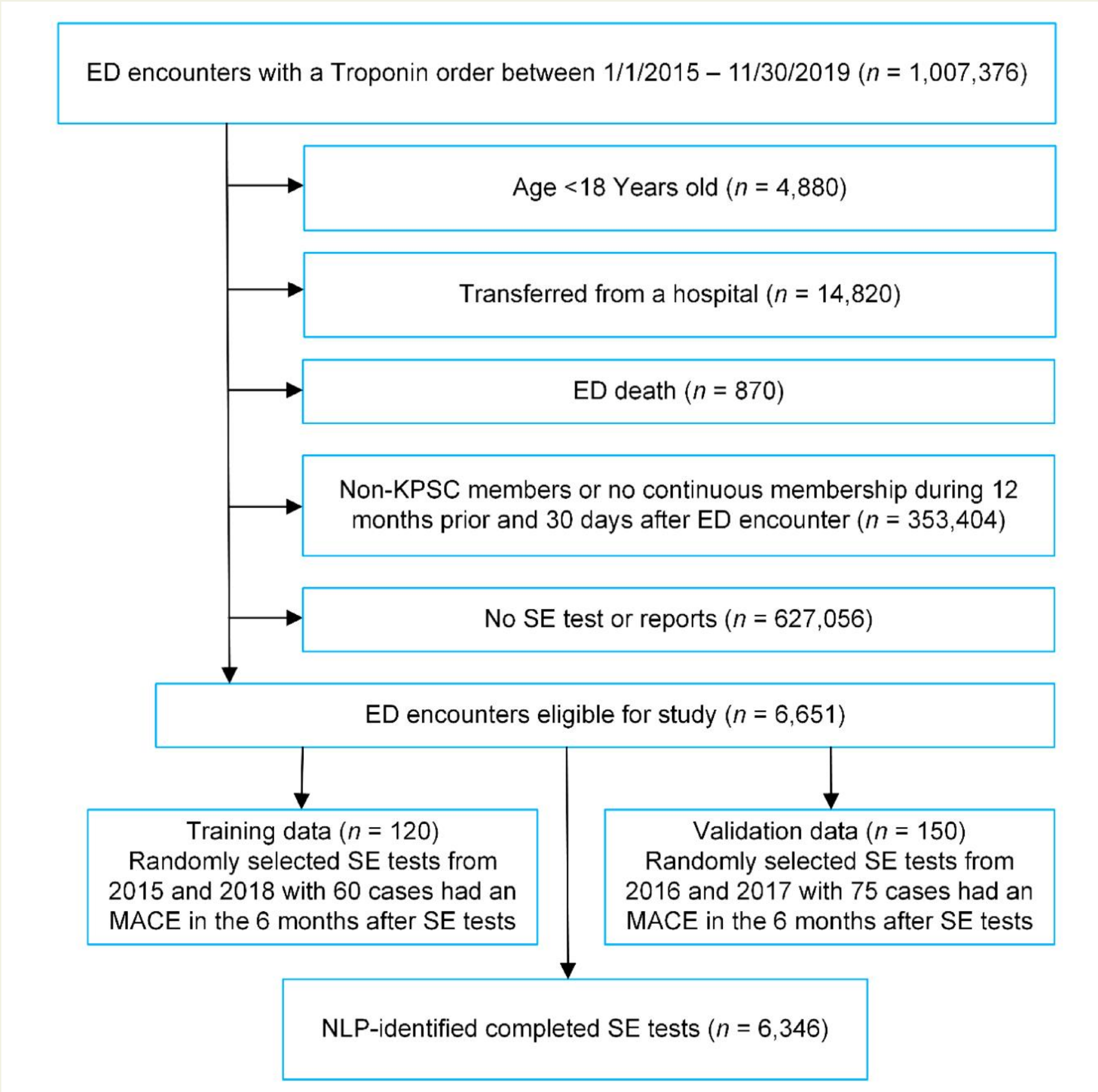


Figure 1 Flowchart for selecting study participants.

Construct validity of natural language processing algorithm

Construct validity is an evaluation of a programme's or measures' overall validity or generalization. It assesses how well a measure is supported by empirical evidence and theoretical patterns. Construct validity was examined in terms of measurement correlations with other variables, as well as its ability to represent characteristics of the condition under investigation.[23] We applied the NLP algorithms to the study cohort and compared the patient characteristics (including HEART and Elixhauser scores) among the different SE results. We treat the SE result as a nominal variable rather than an ordinal variable. We compared the 30-day post-SE cardiac outcomes by the SE results. The descriptive cardiac outcomes included acute myocardial infarction (AMI), cardiac mortality, all-cause mortalities, and MACE rates (a composite of death, AMI, and coronary revascularization). We calculated *P*-values using the χ^2 or the Fisher exact test for all the categorical variables and the Wilcoxon test for all the continuous variables. The significance threshold was set at 0.05. SAS version 9.4 (SAS Institute, Cary, NC, USA) was used for data analysis.

Results

Out of 6651 SE reports in the study period, NLP identified 6346 (95.4%) reports with interpretable overall results (*Table 1*). The

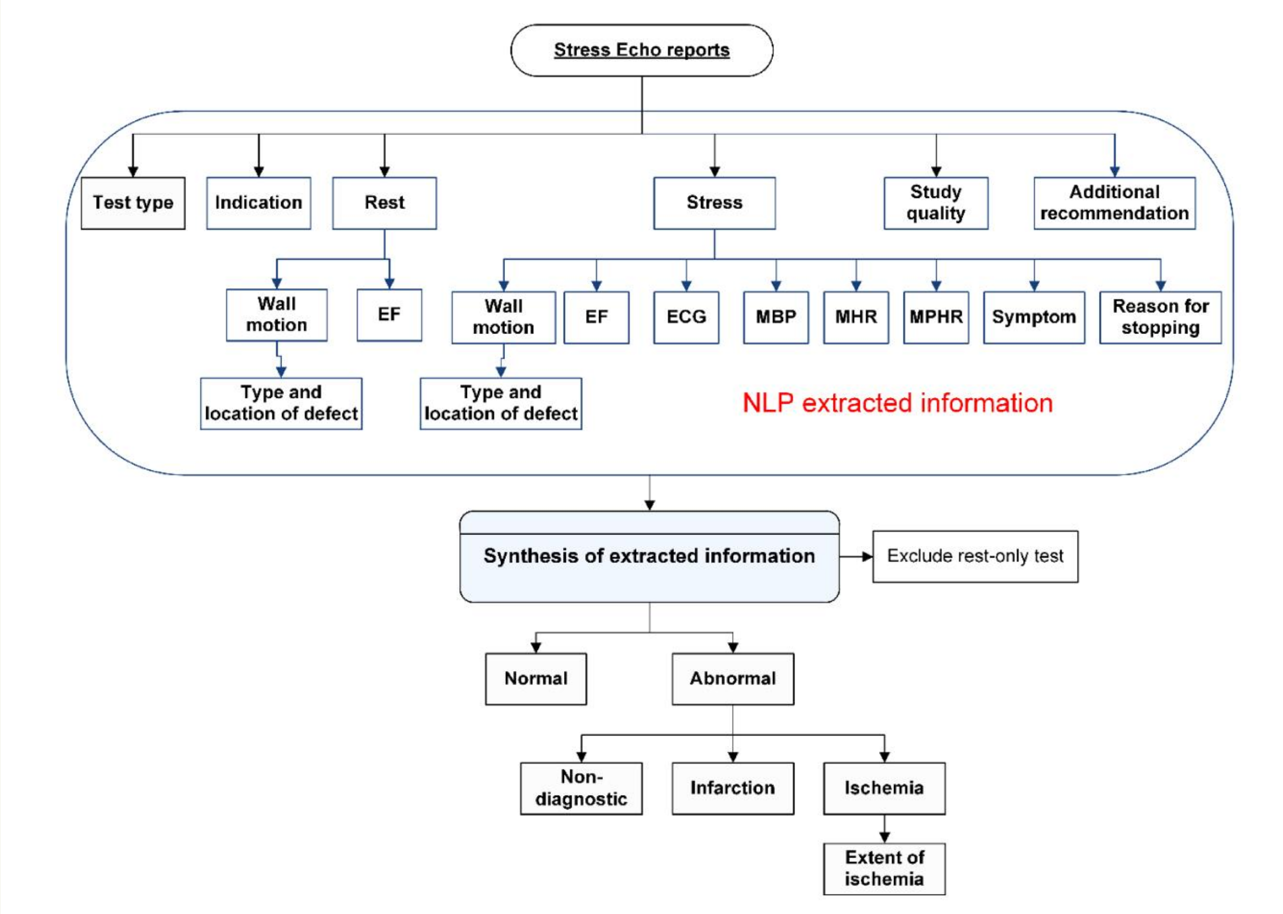


Figure 2 Diagram illustrates the natural language processing process on stress echocardiography reports. Natural language processing extracted commonly available information from the stress echocardiography reports. The extracted information was used to derive the final set of variables based on the clinical logic.

patients' mean age was 63 ± 12 years; 54% were women, 59% were White, and 32% were Hispanic. Over 46% were obese, 42% had a smoking history, and 44% had a family history of CAD. Among the individuals with a HEART score, 15.8, 77.0, and 7.2% had low-, moderate-, and high-risk HEART scores, respectively. The majority (99.5%) of these patients had a troponin level ≤0.5 ng/mL and 83% had an undetectable troponin (<0.02 ng/mL). The numbers of ischaemia, infarction, non-diagnostic, and normal SE results as identified by NLP were 444 (7%), 367 (5.8%), 605 (9.5%), and 4 930 (77.7%), respectively.

Criterion validity of natural language processing algorithm

The final reference standard includes 140 reports after excluding 10 incomplete SE reports. We performed the criterion validity based on these 140 reports. The cardiologists had good agreements on most of the variables in the reference standard (*Table 2*). For the overall result, the Kappa and ICC results were 0.83 and 0.89, respectively. They disagreed more on ischaemic extent (Kappa 0.65, ICC 0.85), rest EF (Kappa 0.75; ICC 0.94), and stress EF (Kappa 0.76; ICC 0.73).

Compared with the reference standard created by the cardiologists, NLP achieved 98.6% specificity and NPV, 95.7% sensitivity, PPV, and *F*-score on SE results using micro-averaged evaluation metrics (*Table 3*). Natural language processing achieved 100% sensitivity, 99.2% specificity, 96.9% PPV, 100% NPV, and 98.2% *F*-score on identifying ischaemia cases. Natural language processing had lower sensitivity, PPV, and *F*-score (85.7%) for infarction. For ischaemia extent, the micro-averaged scores were 98.6% (99.3%) for sensitivity, 100% (99.6%) for specificity, 99.3% (99.6%) for NPV, and 100% (99.3%) for PPV, and 99.3% (99.3%) for *F*-score. Natural language processing had perfect scores on the stress EF categories. For variables not included in the reference standard but validated in our prior studies,[17,18] we also performed manual quality checks.

Construct validity of natural language processing algorithm

We performed the construct validity on the complete set of 6346 reports abstracted by NLP. Exercise was the most common stressor accounting for 3793 (59.8%) of the studies. Chest pain or evaluation for ischaemia is the leading indication for the tests

Table 1 Patient characteristics by natural language processing-identified stress echocardiography results

Characteristic	Ischaemia	Infarction	Non-diagnostics	Normal	*P*-value	Total
n (%)	444 (7)	367 (5.8)	605 (9.5)	4930 (77.7)		6346 (100)
Age	64.5 ± 10.9	65.2 ± 12.7	62.3 ± 12.1	62.5 ± 12.5	<0.0001	62.8 ± 12.4
Female	166 (37.4)	144 (39.2)	343 (56.7)	2773 (56.2)	<0.0001	3426 (54)
Hispanic	110 (24.8)	110 (30)	194 (32.1)	1583 (32.1)	0.02	1997 (31.5)
Race						
White	291 (65.5)	225 (61.3)	347 (57.4)	2872 (58.3)	0.01	3735 (58.9)
Black	39 (8.8)	56 (15.3)	91 (15)	654 (13.3)		840 (13.2)
Asian	42 (9.5)	20 (5.4)	47 (7.8)	438 (8.9)		547 (8.6)
Alaska Native/Pacific Islander	11 (2.5)	3 (0.8)	8 (1.3)	98 (2)		120 (1.9)
Others	61 (13.7)	63 (17.2)	112 (18.5)	868 (17.6)		1104 (17.4)
Body mass index, kg/m^{2}[a]						
<18	2 (0.5)	1 (0.3)	3 (0.5)	31 (0.6)	0.01	37 (0.6)
≥18 and <25	84 (18.9)	69 (18.8)	103 (17)	915 (18.6)		1171 (18.5)
≥25 and <30	159 (35.8)	126 (34.3)	177 (29.3)	1639 (33.2)		2101 (33.1)
≥30 and <35	112 (25.2)	99 (27)	150 (24.8)	1234 (25)		1595 (25.1)
≥35	74 (16.7)	69 (18.8)	167 (27.6)	1020 (20.7)		1330 (21)
Missing	13 (2.9)	3 (0.8)	5 (0.8)	91 (1.8)		112 (1.8)
Smoking behaviour						
Current/passive	36 (8.1)	31 (8.4)	69 (11.4)	334 (6.8)	<0.0001	470 (7.4)
Former	187 (42.1)	159 (43.3)	202 (33.4)	1626 (33)		2174 (34.3)
Never	218 (49.1)	174 (47.4)	332 (54.9)	2893 (58.7)		3617 (57)
Missing	3 (0.7)	3 (0.8)	2 (0.3)	77 (1.6)		85 (1.3)
Family history of CAD	194 (43.7)	158 (43.1)	301 (49.8)	2113 (42.9)	0.02	2766 (43.6)
Elixhauser score	4.1 ± 2.8	5.5 ± 3.1	5.1 ± 3.1	3.8 ± 2.7	<0.0001	4.1 ± 2.8
Comorbidities						
Atrial fibrillation	72 (16.2)	99 (27)	127 (21)	711 (14.4)	<0.0001	1009 (15.9)
CHF	54 (12.2)	98 (26.7)	66 (10.9)	209 (4.2)	<0.0001	427 (6.7)
CAD	213 (48)	197 (53.7)	186 (30.7)	962 (19.5)	<0.0001	1558 (24.6)
Diabetes	182 (41)	146 (39.8)	276 (45.6)	1605 (32.6)	<0.0001	2209 (34.8)
Essential hypertension	320 (72.1)	292 (79.6)	471 (77.9)	3213 (65.2)	<0.0001	4296 (67.7)
Lipid disorder[b]	354 (79.7)	294 (80.1)	480 (79.3)	3564 (72.3)	<0.0001	4692 (73.9)
Renal insufficiency	99 (22.3)	119 (32.4)	155 (25.6)	796 (16.1)	<0.0001	1169 (18.4)
Stroke	15 (3.4)	20 (5.4)	26 (4.3)	139 (2.8)	0.0128	200 (3.2)
Medications[c]						
ACEi/ARB	277 (62.4)	232 (63.2)	307 (50.7)	2228 (45.2)	<0.0001	3044 (48)
Aldosterone	13 (2.9)	13 (3.5)	22 (3.6)	88 (1.8)	0.003	136 (2.1)
Beta blocker	331 (74.5)	224 (61)	323 (53.4)	1568 (31.8)	<0.0001	2446 (38.5)
Calcium channel blockers	93 (20.9)	89 (24.3)	159 (26.3)	985 (20)	0.0014	1326 (20.9)
Diuretics	157 (35.4)	135 (36.8)	238 (39.3)	1443 (29.3)	<0.0001	1973 (31.1)
Vasodilators	26 (5.9)	19 (5.2)	60 (9.9)	192 (3.9)	<0.0001	297 (4.7)
Troponin, ng/mL	0.1 ± 0.22	0.0 + 0.04	0.0 ± 0.05	0.0 ± 0.08	<0.0001	0.0 ± 0.10
HEART score	5.2 ± 1.3	5.2 ± 1.3	4.8 ± 1.3	4.6 ± 1.3	<0.0001	4.7 ± 1.3
Low (0–3)	11 (2.5)	13 (3.5)	34 (5.6)	362 (7.3)	<0.0001	420 (6.6)
Moderate (4–6)	132 (29.7)	120 (32.7)	212 (35)	1585 (32.2)		2049 (32.3)
High (≥7)	26 (5.9)	21 (5.7)	26 (4.3)	120 (2.4)		193 (3)
Missing	275 (61.9)	213 (58)	333 (55)	2863 (58.1)		3684 (58.1)

Values are mean ± SD or *n* (%) unless otherwise indicated. We calculated the *P*-values using the χ^2 test for categorical variables and ANOVA for numerical variables. HEART indicates history, ECG, age, risk factors, and troponin. ACEi, angiotensin-converting enzyme inhibitor; ARB, angiotensin II receptor blockers; CHF, congestive heart failure; CAD, coronary artery disease; NLP, natural language processing.

[a]BMI: the last measure before the ED encounter.

[b]Dyslipidaemia/Hyperlipidaemia.

[c]Medication usage in the 90 days before the emergency department visits.

Table 2 Comparison of agreements between reviewers on abstracting the stress echo reports in the reference standard dataset

Variables	Between two reviewers	
	ICC[a] (95% CI)	Kappa[b] (95% CI)
Test type	0.89 (0.86, 0.92)	0.89 (0.79–0.98)
Echo result[c]	0.89 (0.84–0.82)	0.83 (0.74–0.91)
Ischaemic extent	0.85 (0.80–0.89)	0.65 (0.46–0.83)
Reached MPHR	0.94 (0.92–0.96)	0.90 (0.84–0.97)
Rest EF[d]	0.94 (0.92–0.95)	0.75 (0.63–0.87)
Stress EF[d]	0.73 (0.65–0.80)	0.76 (0.65–0.86)

For Kappa, a value of 0.6–0.79 was considered a moderate agreement; 0.8–0.9 a strong agreement; above 0.9 an almost perfect agreement.[15] For ICC, a coefficient <0.4 was considered poor agreement; 0.4–0.59 was considered fair; 0.6–0.74 was considered good; and >0.75 was considered excellent.[24] CI, confidence interval; EF, ejection fraction; ICC, intraclass correlation; Kappa, Cohen's Kappa; NLP, natural language processing.
[a]Intraclass correlation coefficient, ICC (2,1).
[b]Linear weighted Cohen's Kappa.
[c]Echo result has four possible categories of value: ischaemia, infarction, non-diagnostics, and normal.
[d]Ejection fraction value was converted into categorical values of normal (≥50), and abnormal (<50).

Table 3 Comparison of natural language processing to the reference standard ($n = 140$) on identifying the stress echocardiography results

Confusion matrix	NLP				
Reference standard	Normal	Non-diagnostic	Infarction	Ischaemia	Total
Normal	86	1	1	1	89
Non-diagnostic	2	14	0	0	16
Infarction	1	0	6	0	7
Ischaemia	0	0	0	28	28
Total	89	15	7	29	140

SE result	TP	TN	FN	FP	Sensitivity	Specificity	PPV	NPV	*F*-score
					Accuracy measurements (95% CI)				
Normal[a]	86	48	3	3	96.6 (90.5–99.3)	94.1 (83.8–98.8)	96.6 (90.5–98.6)	94.1 (84.0–98.0)	96.6 (93.8–99.0)
Non-diagnostic[a]	14	123	2	1	87.5 (61.7–98.5)	99.2 (95.6–100)	93.3 (66.3–99.0)	98.4 (94.4–99.6)	90.3 (75.0–100)
Infarction[a]	6	132	1	1	85.7 (42.1–99.6)	99.3 (95.9–100)	85.7 (45.4–977)	99.3 (95.6–99.9)	85.7 (50.0–100)
Ischaemia[a]	28	111	0	1	100 (87.7–100)	99.1 (95.1–100)	96.6 (79.9–99.5)	100	98.2 (93.6–100)
Micro-averaged[b]	134	414	6	6	95.7 (90.9–98.4)	98.6 (96.9–99.5)	95.7 (91.0–98.0)	98.6 (96.9–99.3)	95.7 (92.4–99.1)

CI, confidence interval; SE, myocardial perfusion imaging; NLP, natural language processing; FN, false negative; FP, false positive; TN, true negative; TP, true positive; NPV, negative predictive value; PPV, positive predictive value.
[a]For evaluation purposes, we dichotomized the SE result (e.g. infarction vs. non-infarction) in the confusion matrix to calculate the counts of TP, TN, FN, FP; and derive the performance metrics for each class.
[b]The SE result was evaluated using micro-averaging metrics, which were calculated based on the summarized counts of TP, TN, FN, and FP.

(*Table 4*). There were no statistically significant associations between test indications and SE results. Test reports indicated technical difficulty in 17% of cases; additional testing was recommended for 3.5% of the tests. The proportions of tests with test variables documented in the reports were 98.6% (% of MPHR), 96.5% (maximum heart rate), 95.7% (resting EF), 88.6% (symptom), 84.1% (MRPP), 83.9% (maximum BP), 78.4% (reason for test termination), 65.4% (stress EF).

This study showed an excellent near-term prognostic value of SE (*Figure 3*). Using NLP-abstracted SE summary data, SE demonstrated a good ability to identify individuals at short-term cardiac risk. With worsening SE anomalies, there were significantly higher cardiac event rates. The study cohort's overall 30-day incident rates were 0.8% for death/AMI and 2% for MACE (*Table 5*). The rates of 30-day death/AMI and MACE increased by SE results from normal (0.3 and 0.5%) to non-diagnostic (1.4 and 2.5%), infarction (0.8 and 1.5%),

Table 4 Stress echocardiography variables stratified by the final test results based on the natural language processing algorithm

Variable	Ischaemia	Infarction	Non-diagnostics	Normal	Total
N (%)	444 (7.0)	367 (5.8)	605 (9.5)	4930 (77.7)	6346 (100)
Test type					
Exercise	303 (68.2)	168 (45.8)	307 (50.7)	3015 (61.2)	3793 (59.8)
Dobutamine	141 (31.8)	199 (54.2)	298 (49.3)	1915 (38.8)	2553 (40.2)
Days between ED visit and SE	1.0 ± 1.5	1.0 ± 0.7	0.9 ± 0.9	0.8 ± 0.9	0.9 ± 0.9
Technically difficult	114 (25.7)	72 (19.6)	87 (14.4)	804 (16.3)	1077 (17.0)
Recommend additional test	17 (3.8)	26 (7.1)	113 (18.7)	68 (1.4)	224 (3.5)
Maximum BP	368 (82.9)	307 (83.7)	531 (87.8)	4116 (83.5)	5322 (83.9)
Mean ± SD	174.2 ± 26.6	167.4 ± 27.6	168.3 ± 28.9	174.3 ± 26.0	173.3 ± 26.5
Maximum HR	430 (96.8)	358 (97.5)	586 (96.9)	4747 (96.3)	6121 (96.5)
Mean ± SD	142.8 ± 20.9	139.8 ± 18.8	118.8 ± 20.1	149.5 ± 16.2	145.5 ± 19.4
% of MPHR	439 (98.9)	360 (98.1)	599 (99.0)	4856 (98.5)	6254 (98.6)
Mean ± SD	91.6 ± 13.1	89.9 ± 10.5	74.9 ± 11.3	94.7 ± 7.8	92.3 ± 10.6
MPHR (≥85%)[a]	352 (80.2)	295 (81.9)	40 (6.7)	4855 (100)	5542 (88.6)
MRPP	368 (82.9)	309 (84.2)	532 (87.9)	4125 (83.7)	5334 (84.1)
Mean ± SD	24 895 ± 5559	23 434 ± 5548	19 910 ± 4990	26 168 ± 5104	25 298 ± 5497
MRPP (≥25 000)[a]	188 (51.1)	116 (37.5)	78 (14.7)	2381 (57.7)	2763 (51.8)
EF (rest)					
Abnormal	47 (10.6)	115 (31.3)	0 (0)	1 (0)	163 (2.6)
Normal	368 (82.9)	246 (67.0)	564 (93.2)	4733 (96.0)	5911 (93.1)
Missing	29 (6.5)	6 (1.6)	41 (6.8)	196 (4)	272 (4.3)
EF (stress)					
Abnormal	51 (11.5)	30 (8.2)	4 (0.7)	8 (0.2)	93 (1.5)
Normal	173 (39.0)	222 (60.5)	381 (63)	3282 (66.6)	4058 (63.9)
Missing	220 (49.5)	115 (31.3)	220 (36.4)	1640 (33.3)	2195 (34.6)
Symptom					
Abnormal	188 (42.3)	77 (21.0)	200 (33.1)	694 (14.1)	1159 (18.3)
Normal	169 (38.1)	260 (70.8)	365 (60.3)	3671 (74.5)	4465 (70.4)
Missing	87 (19.6)	30 (8.2)	40 (6.6)	565 (11.5)	722 (11.4)
Cardiac	66 (14.9)	20 (5.4)	104 (17.2)	121 (2.5)	311 (4.9)
Noncardiac	52 (11.7)	30 (8.2)	115 (19.0)	275 (5.6)	472 (7.4)
Endpoint	243 (54.7)	217 (59.1)	267 (44.1)	3467 (70.3)	4194 (66.1)
Missing	83 (18.7)	100 (27.2)	119 (19.7)	1067 (21.6)	1369 (21.6)

Not all NLP-abstracted variables were listed in this table.
P-value was <0.0001 for all variables listed, except for 'Days between ED visit and SE' ($P = 0.007$).
Ejection fraction was grouped into two categories based on the qualitative or quantitative documentation:
• Abnormal: qualitative (mild, moderate, or severe abnormal) or quantitative (<50%).
• Normal: qualitative (normal, hyperdynamic) or quantitative (≥50%).
BP, blood pressure; ED, emergency department; EF, ejection fraction; HR, heart rate; MPHR, maximum predicted heart rate; MRPP, maximum rate pressure product; NLP, natural language processing.
[a]The denominators for MPHR (≥85%) and MRPP (≥25 000) were the numbers of reports with MPHR or MRPP.

and ischaemia (5.0 and 18.5%). Patients with ischaemia had a 17-fold increased 30-day death/AMI rates compared with patients with normal SE.

Discussion

This study showed that NLP may be reliably used to classify SE findings and abstract variables that are often used in stress testing score models.[25] This is the first study we are aware of that uses a computer-based method to abstract SE findings. We found that NLP could provide a coherent summary interpretation by synthesizing the data elements presented in the reports. Compared with the cardiologists' adjudicated reference standard, the NLP algorithm had high accuracy for criterion validity. We also performed construct validity of the NLP algorithm on the SE tests' near-term cardiac outcomes. Ischaemia by SE was an independent predictor of future cardiac outcomes as reported by previous studies.[5,26,27] Our NLP-derived SE final results had good differentiating power in identifying patients at short-term cardiac risk, which is consistent with

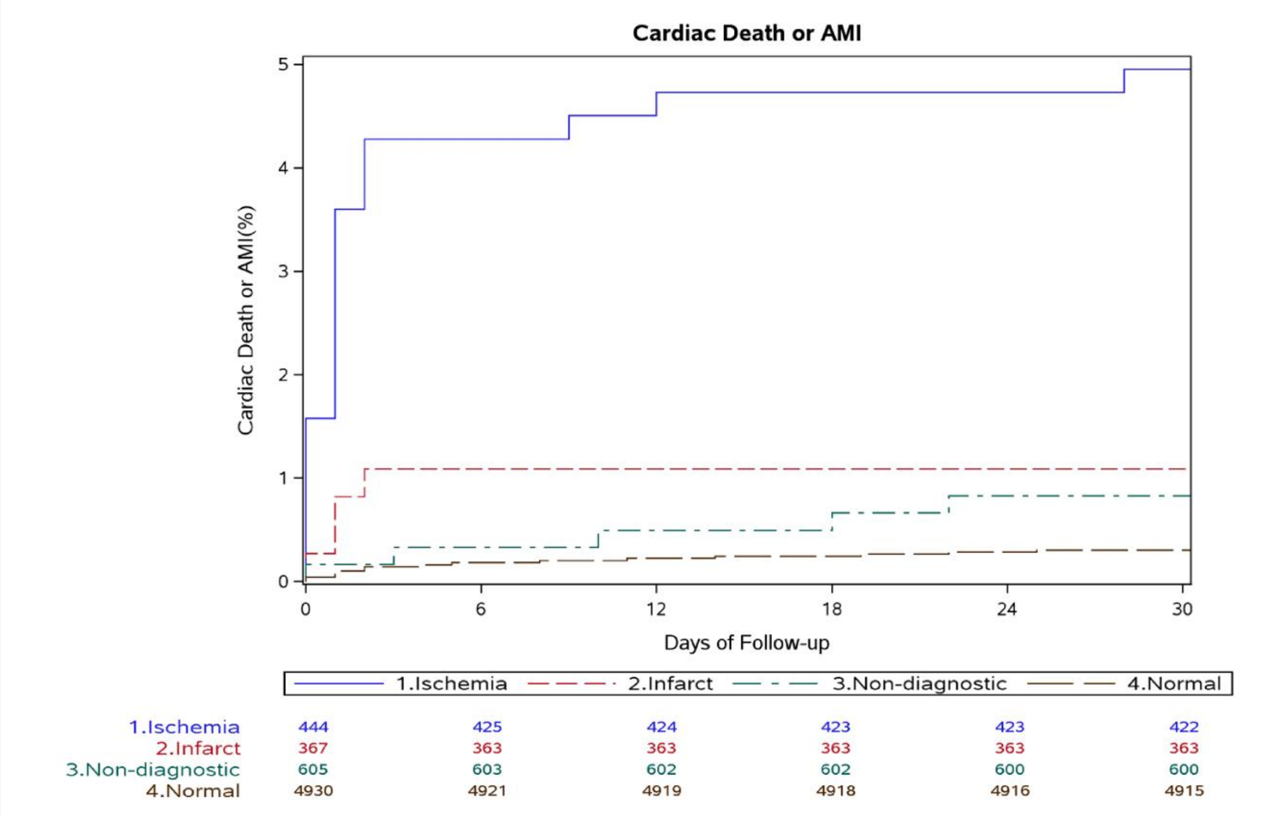


Figure 3 Cumulative event rate for cardiac death or acute myocardial infarction for patients who underwent stress echocardiography after emergency department visits ($n = 6346$). The horizontal line indicates the number of days since the index (emergency department) visit in the study cohort who had complete stress echocardiography reports ($n = 6346$). The χ^2 value from the log-rank test is 125.8 with a P-value < 0.0001.

prior results. We intend to publish a more comprehensive construct validation study in the future, comparing patients' characteristics, SE test variables, test modalities, and cardiac outcomes stratified by NLP's results.

Quality of stress echo reports

While clinical societies have promoted quality improvement in cardiovascular imaging reporting for more than a decade,[9,28] the majority of echo reports is still in semi-structured or unstructured formats with many variations between different institutions.[29] Our study found that, even in the same institution, there were substantial differences in the format and quality of SE reports. Stress echocardiography reports frequently contained ambiguous and hedging language, as illustrated by the sample reports (Supplementary material online, *Results S1–S4*), making proper interpretation challenging.

In the year 2020, the American Society of Echocardiography published guidelines for SE performance, interpretation, and application in ischaemic heart disease.[19] The guidelines made no mention of structured reporting, but they did provide recommendations for data items in SE reports. In this study, the data completeness of SE reports varies by the variables, from 65% to near 100%. In particular, only 65% of reports documented stress EF, which was a data element in the guidelines.[19]

Some variables were often implicitly stated. For instance, the new guidelines stated that interpretation must summarize the extent, severity, location of wall motion abnormalities, and their correlation with coronary anatomy.[19] However, in the community settings of this large healthcare system, most SE reports did not include explicit information about the extent and severity of wall motion abnormalities, which were often available in the MPI reports in our prior study.[18] The NLP algorithm derived the ischaemic extent based on the involved segments, which were often described using vague, inconsistent, or non-standard terms (Supplementary material online, *Result S1*).

The abnormalities in the rest and stress tests were sometimes mentioned separately, with no clear final interpretation. In this study, only 175 out of 475 infarction cases were explicitly stated in the reports (Supplementary material online, *Result S2*). Critical information was sometimes documented in the 'Findings' section rather than the 'Conclusion' section, which is often the only section read by the referring providers.

Structured reporting may reduce but not eliminate the quality issues in the SE reports.[30]

Integrating automatic tools such as the NLP programme into clinical care could help identify reporting errors and quality gaps, which will help improve patient outcomes.[31] The referring providers and primary care physicians perform risk assessments[2] and relay results to patients after reading the SE reports.[19,30] The echocardiography

Table 5 The 30-day cardiac outcomes stratified by natural language processing-identified stress echocardiography results

	Ischaemia	Infarction	Non-diagnostics	Normal	Total
N (%)	444 (7.0)	367 (5.8)	605 (9.5)	4930 (77.7)	6346 (100)
Unstable angina	41 (9.2)	2 (0.5)	11 (1.8)	25 (0.5)	79 (1.2)
MACE	82 (18.5)	9 (2.5)	9 (1.5)	25 (0.5)	125 (2)
AMI or all-cause death	22 (5.0)	5 (1.4)	5 (0.8)	17 (0.3)	49 (0.8)
AMI	22 (5)	4 (1.1)	5 (0.8)	15 (0.3)	46 (0.7)
All-cause death	0 (0)	1 (0.3)	0 (0)	2 (0)	3 (0)
AMI or cardiac death	22 (5.0)	5 (1.4)	5 (0.8)	17 (0.3)	49 (0.8)
Cardiac death	0 (0)	1 (0.3)	0 (0)	1 (0)	2 (0)
Revascularization	73 (16.4)	4 (1.1)	6 (1)	14 (0.3)	97 (1.5)
CABG	50 (11.3)	2 (0.5)	4 (0.7)	2 (0)	58 (0.9)
PCI	25 (5.6)	2 (0.5)	2 (0.3)	12 (0.2)	41 (0.6)

Data were presented as *n* (%). *P*-values were calculated using Fisher's Exact Test. All *P*-values were <0.0001 for variables listed in this table, except for the 30-day death ($P = 0.23$). Revascularization includes CABG and PCI. AMI, acute myocardial infarction; CABG, coronary artery bypass grafting; MACE, major adverse cardiac events which include AMI, death, and coronary revascularization; PCI, percutaneous coronary intervention.

reports are not always well understood by non-cardiology providers.[32] Moreover, patients are also increasingly accessing their EMR and actively participating in treatment decisions. This NLP technique could be used to provide test summaries in languages that general clinicians and patients can understand.

Potential usages of the natural language processing algorithm

Traditional SE reports, which are dictated or typed, are highly variable depending on physician (cardiologist) preferences and practices, making it challenging for referring physicians (e.g. non-cardiologists), researchers, or automated abstraction to understand the findings. Moreover, population-based studies of SE would typically be difficult to carry out due to the time and resources needed for manual chart review. This tool could create new opportunities for research and care improvement in several ways, including its ability to dramatically reduce the time required for chart review, as well as increase accuracy through a more consistent approach to abstracting SE reports. Thus, this NLP algorithm could support large-scale research studies using real-world data, such as the cost-effectiveness and appropriate population of SE tests.

Natural language processing might also be used to create summaries for SE reporting that are easier for patients and referring physicians to understand. The referring and primary care physicians could benefit from important findings highlighted by NLP. A machine learning model could use NLP-derived data to aid in identifying and reducing future risks. We previously utilized NLP to identify patients with lower EF values from echocardiography records as part of a system-wide tracking and alarm programme that enhanced patient safety and filled therapy gaps.[33] Similarly, this technology could lead to better medical care.

Study strengths and limitations

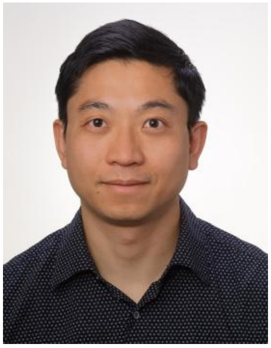
We validated our algorithm on a large and diverse population within an integrated care system with a comprehensive EMR. Furthermore, our prepaid health plan minimized the racial disparity in seeking medical care. The prepaid model, on the other hand, has no financial influence on the providers' choices of noninvasive tests. Moreover, few studies have looked at the predictive value of SE in near-term cardiac events in patients referred from ED. Despite the low event rates, we were able to assess near-term outcomes due to the size of our study cohort.

Our study has some limitations. Stress echocardiography results were based on the reading physicians' interpretations of the echo images. Variations in the accuracy of the test interpretation are expected among physicians.[19,20] We did not have the resources to validate the written SE reports by re-examining the SE images. However, the NLP-derived data could be used as the labelled data in developing an automated image recognition system based on machine learning or deep learning. Furthermore, we limited our analyses using the overall interpretation since it is often the only information used in clinical decision-making by the referring providers. The other variables extracted by NLP could augment the SE results for a better outcome prediction. In the future, we may use statistical and machine learning methods to build a predictive model that offers better prognostic value than the overall interpretation. Moreover, while SE is used to assess a wide range of cardiac conditions, this study used the NLP method to identify IHD in ED populations. In the future, we may extend this NLP method to identify other cardiac conditions from the SE reports. Finally, although this study includes data from 15 medical centres, the language and style of reporting can vary between institutions. Institutional changes to reader interpretation and reporting could also occur. When a performance decline is noted, the algorithm may need to be retrained.

Conclusions

We developed and validated an automated NLP algorithm to abstract the conventional SE reports with high accuracy. It had near-perfect accuracy in identifying ischaemia results, which is the most critical data in the SE report.

Lead author biography



Chengyi Zheng is a principal data scientist in Kaiser Permanente Southern California, USA. He earned his PhD in computer science with an artificial intelligence specialization from Oregon Health and Science University. He has 19 years' experience working on machine learning and natural language processing projects for the pharmaceutical, healthcare, and medical device industries. In 2016 and 2021, he won first prize in the Kaiser Permanente data science competition and NIST's TREC Clinical Trial Task, respectively. He has worked on various cardiovascular research initiatives based on artificial intelligence.

Supplementary material

Supplementary material is available at European Heart Journal—Digital Health.

Funding

This work was supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under Award Number R01HL134647. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of interest: Author, B.C.S., was a consultant for Medtronic. The remaining authors have no conflicts of interest to report.

Data availability

The data underlying this article cannot be shared publicly due to HIPPA and institutional restrictions.

References

1. Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, Barengo NC, Beaton AZ, Benjamin EJ, Benziger CP, Bonny A, Brauer M, Brodmann M, Cahill TJ, Carapetis J, Catapano AL, Chugh SS, Cooper LT, Coresh J, Criqui M, DeCleene N, Eagle KA, Emmons-Bell S, Feigin VL, Fernández-Solà J, Fowkes G, Gakidou E, Grundy SM, He FJ, Howard G, Hu F, Inker L, Karthikeyan G, Kassebaum N, Koroshetz W, Lavie C, Lloyd-Jones D, Lu HS, Mirijello A, Temesgen AM, Mokdad A, Moran AE, Muntner P, Narula J, Neal B, Ntsekhe M, Moraes de Oliveira G, Otto C, Owolabi M, Pratt M, Rajagopalan S, Reitsma M, Ribeiro ALP, Rigotti N, Rodgers A, Sable C, Shakil S, Sliwa-Hahnle K, Stark B, Sundström J, Timpel P, Tleyjeh IM, Valgimigli M, Vos T, Whelton PK, Yacoub M, Zuhlke L, Murray C, Fuster V, GBD-NHLBI-JACC Global Burden of Cardiovascular Diseases Writing Group. Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019: update from the GBD 2019 study. *J Am Coll Cardiol* 2020;**76**:2982–3021.
2. Gulati M, Levy PD, Mukherjee D, Amsterdam E, Bhatt DL, Birtcher KK, Blankstein R, Boyd J, Bullock-Palmer RP, Conejo T, Diercks DB, Gentile F, Greenwood JP, Hess EP, Hollenberg SM, Jaber WA, Jneid H, Joglar JA, Morrow DA, O'Connor RE, Ross MA, Shaw LJ. 2021 AHA/ACC/ASE/CHEST/SAEM/SCCT/SCMR Guideline for the Evaluation and Diagnosis of Chest Pain: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Circulation* 2021;**144**:e368–e454.
3. Foy AJ, Liu G, Davidson WR, 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.
4. Siontis GC, Mavridis D, Greenwood JP, Coles B, Nikolakopoulou A, Jüni P, Salanti G, Windecker S. Outcomes of non-invasive diagnostic modalities for the detection of coronary artery disease: network meta-analysis of diagnostic randomised controlled trials. *BMJ* 2018;**360**:k504.
5. Metz LD, Beattie M, Hom R, Redberg RF, Grady D, Fleischmann KE. The prognostic value of normal exercise myocardial perfusion imaging and exercise echocardiography: a meta-analysis. *J Am Coll Cardiol* 2007;**49**:227–237.
6. Knuuti J, Wijns W, Saraste A, Capodanno D, Barbato E, Funck-Brentano C, Prescott E, Storey RF, Deaton C, Cuisset T, Agewall S, Dickstein K, Edvardsen T, Escaned J, Gersh BJ, Svitil P, Gilard M, Hasdai D, Hatala R, Mahfoud F, Masip J, Muneretto C, Valgimigli M, Achenbach S, Bax JJ, Neumann F-J, Sechtem U, Banning AP, Bonaros N, Bueno H, Bugiardini R, Chieffo A, Crea F, Czerny M, Delgado V, Dendale P, Flachskampf FA, Gohlke H, Grove EL, James S, Katritsis D, Landmesser U, Lettino M, Matter CM, Nathoe H, Niessner A, Patrono C, Petronio AS, Pettersen SE, Piccolo R, Piepoli MF, Popescu BA, Räber L, Richter DJ, Roffi M, Roithinger FX, Shlyakhto E, Sibbing D, Silber S, Simpson IA, Sousa-Uva M, Vardas P, Witkowski A, Zamorano JL, Achenbach S, Agewall S, Barbato E, Bax JJ, Capodanno D, Cuisset T, Deaton C, Dickstein K, Edvardsen T, Escaned J, Funck-Brentano C, Gersh BJ, Gilard M, Hasdai D, Hatala R, Mahfoud F, Masip J, Muneretto C, Prescott E, Saraste A, Storey RF, Svitil P, Valgimigli M, Windecker S, Aboyans V, Baigent C, Collet J-P, Dean V, Delgado V, Fitzsimons D, Gale CP, Grobbee D, Halvorsen S, Hindricks G, Iung B, Jüni P, Katus HA, Landmesser U, Leclercq C, Lettino M, Lewis BS, Merkely B, Mueller C, Petersen S, Petronio AS, Richter DJ, Roffi M, Shlyakhto E, Simpson IA, Sousa-Uva M, Touyz RM, Benkhedda S, Metzler B, Sujayeva V, Cosyns B, Kusljugic Z, Velchev V, Panayi G, Kala P, Haahr-Pedersen SA, Kabil H, Ainla T, Kaukonen T, Cayla G, Pagava Z, Woehrle J, Kanakakis J, Tóth K, Gudnason T, Peace A, Aronson D, Riccio C, Elezi S, Mirrakhimov E, Hansone S, Sarkis A, Babarskiene R, Beissel J, Maempel AJC, Revenco V, de Grooth GJ, Pejkov H, Juliebø V, Lipiec P, Santos J, Chioncel O, Duplyakov D, Bertelli L, Dikic AD, Studenčan M, Bunc M, Alfonso F, Bäck M, Zellweger M, Addad F, Yildirir A, Sirenko Y, Clapp B. 2019 ESC guidelines for the diagnosis and management of chronic coronary syndromes. *Eur Heart J* 2020;**41**:407–477.
7. van Waardhuizen CN, Khanji MY, Genders TSS, Ferket BS, Fleischmann KE, Hunink MGM, Petersen SE. Comparative cost-effectiveness of non-invasive imaging tests in patients presenting with chronic stable chest pain with suspected coronary artery disease: a systematic review. *Eur Heart J Qual Care Clin Outcomes* 2016;**2**:245–260.
8. Amsterdam EA, Wenger NK, Brindis RG, Casey DE, Ganiats TG, Holmes DR, Jaffe AS, Jneid H, Kelly RF, Kontos MC, Levine GN, Liebson PR, Mukherjee D, Peterson ED, Sabatine MS, Smalling RW, Zieman SJ. 2014 AHA/ACC guideline for the management of patients with non-ST-elevation acute coronary syndromes: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* 2014;**130**:e344–426.
9. Gardin JM, Adams DB, Douglas PS, Feigenbaum H, Forst DH, Fraser AG, Grayburn PA, Katz AS, Keller AM, Kerber RE, Khandheria BK, Klein AL, Lang RM, Pierard LA, Quinones MA, Schnittger I. Recommendations for a standardized report for adult transthoracic echocardiography: a report from the American Society of Echocardiography's Nomenclature and Standards Committee and Task Force for a Standardized Echocardiography Report. *J Am Soc Echocardiogr* 2002;**15**:275–290.
10. Dey D, Slomka PJ, Leeson P, Comaniciu D, Shrestha S, Sengupta PP, Marwick TH. Artificial intelligence in cardiovascular imaging: JACC state-of-the-art review. *J Am Coll Cardiol* 2019;**73**:1317–1335.
11. 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.
12. 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.e2.
13. Al'Aref SJ, Anchouche K, Singh G, Slomka PJ, Kolli KK, Kumar A, Pandey M, Maliakal G, van Rosendael AR, Beecy AN, Berman DS, Leipsic J, Nieman K, Andreini D, Pontone G, Schoepf UJ, Shaw LJ, Chang H-J, Narula J, Bax JJ, Guan Y, Min JK. Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. *Eur Heart J* 2019;**40**:1975–1986.
14. Sharp AL, Wu Y-L, Shen E, Redberg R, Lee M-S, 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.
15. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)* 2012;**22**:276–282.
16. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;**86**:420–428.
17. Zheng C, Sun BC, Wu Y, Lee M, Shen E, Redberg RF, Ferencik M, Natsui S, Kawatkar AA, Musgidlok VV, Sharp AL. Automated identification and extraction of exercise treadmill test results. *J Am Heart Assoc* 2020;**9**:e014940.

18. Zheng C, Sun BC, Wu Y-L, Ferencik M, Lee M-S, Redberg RF, Kawatkar AA, Musigdilok VV, Sharp AL. Automated abstraction of myocardial perfusion imaging reports using natural language processing. *J Nucl Cardiol* 2022;**29**:1178–1187.
19. Pellikka PA, Arruda-Olson A, Chaudhry FA, Chen MH, Marshall JE, Porter TR, Sawada SG. Guidelines for performance, interpretation, and application of stress echocardiography in ischemic heart disease: from the American Society of Echocardiography. *J Am Soc Echocardiogr* 2020;**33**:1–41.e8.
20. Marwick TH. Stress echocardiography. *Heart* 2003;**89**:113–118.
21. Shaw LJ, Berman DS, Picard MH, Friedrich MG, Kwong RY, Stone GW, Senior R, Min JK, Hachamovitch R, Scherrer-Crosbie M, Mieres JH, Marwick TH, Phillips LM, Chaudhry FA, Pellikka PA, Slomka P, Arai AE, Iskandrian AE, Bateman TM, Heller GV, Miller TD, Nagel E, Goyal A, Borges-Neto S, Boden WE, Reynolds HR, Hochman JS, Maron DJ, Douglas PS. Comparative definitions for moderate-severe ischemia in stress nuclear, echocardiography, and magnetic resonance imaging. *JACC Cardiovasc Imaging* 2014;**7**:593–604.
22. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Inf Process Manag* 2009;**45**:427–437.
23. Portney LG, Watkins MP. *Foundations of clinical research: applications to practice*. Upper Saddle River, NJ: Pearson/Prentice Hall; 2009.
24. Cicchett, Domenic V. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess. 1994;6(4):284–90.
25. Ashley E, Myers J, Froelicher V. Exercise testing scores as an example of better decisions through science. *Med Sci Sports Exerc* 2002;**34**:1391–1398.
26. Beleslin BD, Ostojic M, Stepanovic J, Djordjevic-Dikic A, Stojkovic S, Nedeljkovic M, Stankovic G, Petrasinovic Z, Gojkovic L, Vasiljevic-Pokrajcic Z. Stress echocardiography in the detection of myocardial ischemia. Head-to-head comparison of exercise, dobutamine, and dipyridamole tests. *Circulation* 1994;**90**:1168–1176.
27. Rallidis L, Cokkinos P, Tousoulis D, Nihoyannopoulos P. Comparison of dobutamine and treadmill exercise echocardiography in inducing ischemia in patients with coronary artery disease. *J Am Coll Cardiol* 1997;**30**:1660–1668.
28. Douglas P, Iskandrian AE, Krumholz HM, Gillam L, Hendel R, Jollis J, Peterson E, Douglas P, Chen J, Gillam L, Hendel R, Jollis J, Iskandrian AE, Krumholz HM, Masoudi F, Mohler E, McNamara RL, Patel MR, Peterson E, Spertus J. Achieving quality in cardiovascular imaging: proceedings from the American College of Cardiology-Duke University Medical Center Think Tank on Quality in Cardiovascular Imaging. *J Am Coll Cardiol* 2006;**48**:2141–2151.
29. Adekkanattu P, Jiang G, Luo Y, Kingsbury PR, Xu Z, Rasmussen LV, Pacheco JA, Kiefer RC, Stone DJ, Brandt PS, Yao L. Evaluating the portability of an NLP system for processing echocardiograms: a retrospective, multi-site observational study. *AMIA Annu Symp Proc* 2019;**2019**:190–199.
30. Eskandari M, Kramer CM, Hecht HS, Jaber WA, Marwick TH. Evidence base for quality control activities in cardiovascular imaging. *JACC Cardiovasc Imaging* 2016;**9**: 294–305.
31. Chandra S, Arling B, Rock J, Spencer KT. Detection of discrepancies in facilitated echocardiographic reporting using a prototype rule generator. *J Am Soc Echocardiogr* 2010;**23**:778–782.
32. Trang A, Kampangkaew J, Fernandes R, Tiwana J, Misra A, Hamzeh I, Blaustein A, Aguilar D, Shah T, Ballantyne C, Quinones M, Nagueh SF, Dokanish H, Virani SS, Deswal A, Kirkpatrick JN, Nambi V. Understanding by general providers of the echocardiogram report. *Am J Cardiol* 2019;**124**:296–302.
33. Danforth KN, Smith AE, Loo RK, Jacobsen SJ, Mittman BS, Kanter MH. Electronic clinical surveillance to improve outpatient care: diverse applications within an integrated delivery system. *EGEMS (Wash DC)* 2014;**2**:1056.