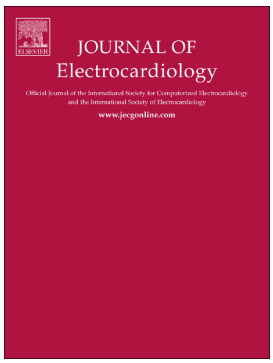

A deep neural network learning algorithm outperforms a conventional algorithm for emergency department electrocardiogram interpretation

Stephen W. Smith, Brooks Walsh, Ken Grauer, Kyuhyun Wang, Jeremy Rapin, Jia Li, William Fennell, Pierre Taboulet

PII: S0022-0736(18)30229-2
DOI: https://doi.org/10.1016/j.jelectrocard.2018.11.013
Reference: YJELC 52763

To appear in: *Journal of Electrocardiology*

Please cite this article as: Stephen W. Smith, Brooks Walsh, Ken Grauer, Kyuhyun Wang, Jeremy Rapin, Jia Li, William Fennell, Pierre Taboulet , A deep neural network learning algorithm outperforms a conventional algorithm for emergency department electrocardiogram interpretation. Yjelc (2018), https://doi.org/10.1016/j.jelectrocard.2018.11.013

A Deep Neural Network Learning Algorithm Outperforms a Conventional Algorithm For Emergency Department Electrocardiogram Interpretation

Stephen W. Smith, MD*¶, Brooks Walsh, MD+, Ken Grauer, MD@, Kyuhyun Wang, MD§, Jeremy Rapin, Ph.D.$, Jia Li $, William Fennell, M.D.£, Pierre Taboulet, M.D.$#

* Department of Emergency Medicine, Hennepin County Medical Center, Minneapolis, Minnesota

¶ University of Minnesota Department of Emergency Medicine

+ Bridgeport Hospital, Bridgeport, Connecticut, USA

@ College of Medicine, University of Florida

£ Department of Cardiology, University College, Cork, Ireland.

§ University of Minnesota, Department of Medicine, Division of Cardiology

Cardiologist, Department of Emergency Medicine, Hôpital Saint Louis, Assistance Publique–Hôpitaux de Paris, Paris, France

$ Cardiologs® Technologies, Paris, France

Word count (Introduction through conclusion, legends, references and appendix)

4715

Funding

All funding was by Cardiologs® technologies.

The first author, Dr. Stephen W. Smith, received no funding and received no financial support of any kind from any source.

Disclosures

Brooks Walsh -- paid by Cardiologs® a set fee for blind interpretation of ECGs

Kyuhyun Wang -- paid by Cardiologs® a set fee for blind interpretation of ECGs

Ken Grauer -- paid by Cardiologs® a set fee for blind interpretation of ECGs

Jia Li -- employed by Cardiologs®, shareholder in Cardiologs®

Jeremy Rapin -- employed by Cardiologs®, shareholder in Cardiologs®

William Fennell -- no disclosures

Pierre Taboulet -- shareholder in Cardiologs®

Stephen Smith -- no disclosures

Address for Correspondence:

Stephen W. Smith

Department of Emergency Medicine

Hennepin County Medical Center

715 Park Ave., ER R-2

Minneapolis, MN, 55415

612-875-4226

smith253@umn.edu

612-904-4241

Abstract (274 words)

Background

Cardiologs® has developed the first electrocardiogram (ECG) algorithm that uses a deep neural network (DNN) for full 12-lead ECG analysis, including rhythm, QRS and ST-T-U waves. We compared the accuracy of the first version of Cardiologs® DNN algorithm to the Mortara / Veritas® conventional algorithm in emergency department (ED) ECGs.

Methods

Individual ECG diagnoses were prospectively mapped to one of 16 pre-specified groups of ECG diagnoses, which were further classified as "major" ECG abnormality or not. Automated interpretations were compared to blinded experts'. The primary outcome was the performance of the algorithms in finding at least one "major" abnormality. The secondary outcome was the proportion of all ECGs for which all groups were identified, with no false negative or false positive groups ("accurate ECG interpretation"). Additionally, we measured sensitivity and positive predictive value (PPV) for any abnormal group.

Results

Cardiologs® vs. Veritas® accuracy for finding a major abnormality was 92.2% vs. 87.2% ($p < 0.0001$), with comparable sensitivity (88.7% vs. 92.0%, $p = 0.086$), improved specificity (94.0% vs. 84.7% , $p < 0.0001$) and improved positive predictive value (PPV 88.2% vs. 75.4%, $p < 0.0001$). Cardiologs® had accurate ECG interpretation for 72.0% (95% CI: 69.6-74.2) of ECGs vs. 59.8% (57.3-62.3) for Veritas® ($P < 0.0001$). Sensitivity for any abnormal group for Cardiologs® and Veritas®, respectively, was 69.6% (95CI 66.7-72.3) vs. 68.3%

(95CI 65.3-71.1) (NS). Positive Predictive Value was 74.0% (71.1-76.7) for Cardiologs® vs. 56.5% (53.7-59.3) for Veritas® (P < 0.0001).

Conclusion

Cardiologs' DNN was more accurate and specific in identifying ECGs with at least one major abnormal group. It had a significantly higher rate of accurate ECG interpretation, with similar sensitivity and higher PPV.

Keywords

Electrocardiography, Computer, Deep Neural Network, Artificial Intelligence, Big Data

Abbreviations

ECG = electrocardiogram

BBB = bundle branch block

STEMI = ST Elevation Myocardial Infarction

Non-STEMI = Non ST Elevation Myocardial Infarction

HR = heart rate

ED = emergency department

DNN = deep neural network

PPV = positive predictive value

NPV = negative predictive value

AV = atrio-ventricular

Introduction

Computer electrocardiogram (ECG) interpretation algorithms aim to improve physician ECG interpretation, reduce medical error, and expedite patient care. Interpretations include both rhythm analysis and QRS-T-U analysis. Rapid interpretation is particularly important in the emergency department (ED), and although critical care and emergency physicians must be experienced in ECG interpretation, improved accuracy of automated interpretations could improve efficiency and patient safety.[1,2] However, computer algorithms have had mediocre performance.[2] Recent algorithms were only approximately 65% sensitive and 90% specific for ST Elevation Myocardial Infarction.[3,4] Furthermore, erroneous automated interpretations are associated with erroneous physician overreads, whereas accurate interpretations are associated with accurate physician overreads.[5-7] When ECGs are over-read by cardiologists, the presence of an automated interpretation results in lower accuracy, as the automated errors are frequently not corrected.[8] Erroneous computerized interpretations of atrial fibrillation, or its absence, have been correlated with erroneous final overreads, which adversely affected management.[7] Thus, an initial correct automated interpretation importantly influences the final interpretation and, consequently, patient management. In spite of the importance and ubiquity of computer ECG interpretation, we are unaware of any publication of a direct comparative evaluation of any of the many algorithms in use; recent reviews made the same observation and also call for action to improve computerized ECG interpretation.[2,9]

Deep Neural Network (DNN) machine learning algorithms use large amounts of data to "train" the computer by labeling each case according to one of many predefined abnormalities, allowing the computer to discern what characteristics of ECGs are associated with any given abnormality. A great advantage to DNN is they keep learning: the quality and accuracy is continuously refined as more data with outcomes or corresponding accurate

interpretations accumulate. Unlike expert humans, who may retire or die, the DNN can continue to learn indefinitely. Eric Topol lists deep neural network learning application to skin cancer, as reported in Nature,[10] as third among the top ten technological advances in medicine in 2017.[11]

DNN have been shown to be accurate for isolated labels of the ECG, such as rhythm diagnosis, in which it exceeded cardiologists' performance for arrhythmia detection.[12] However, Cardiologs Technologies has produced a new DNN algorithm (20 M parameters, 1.6 M neurons, 16 layers) which is not limited to interpretation of any one ECG label, but rather, for the first time, interprets the entire 12-lead ECG, including rhythm and QRS-T-U waves. The Cardiologs' DNN has already been shown to perform better than existing solutions, including Veritas®, at detecting atrial fibrillation.[13,14] The goal of this study is to assess the performance of this algorithm in the context of an ED, and with a larger scope of pathologies.

By the start of this study, this new Cardiologs Technologies DNN, had been trained on approximately 130,000 ECGs which were annotated by expert interpreters, with analysis of both conduction and the QRS-T-U waves. Training ECGs were recorded at multiple institutions, including at Hennepin County Medical Center (HCMC) in Minneapolis, MN, USA, from January 2013 through June 2016. The result is a function which takes an ECG as input and provides the probabilities of the presence of a list of abnormalities as output, producing a full rhythm and QRS-T-U analysis. This computation is performed using a sequence of simple operations such as matrix products. The coefficients of these matrices are however not directly interpretable, such that the internal workings of the algorithm are very difficult to decipher.

We sought to evaluate the performance of the first version of this new algorithm in a context where Cardiologs' solution is typically used: an emergency department, an area where time and clarity is important for optimal decision making and patient management.

In the emergency department, where an ECG is deemed to be necessary, the logical thought process for the reviewer is the following: Is this ECG normal or abnormal? If it is abnormal, is it an emergency? If it is abnormal but not an emergency, is it significant or non-significant? We compared Cardiologs' solution to the widely used Veritas® conventional algorithm used on Mortara® ECG machines, following this logical thought process.

Methods

This was a retrospective study of ECGs of patients presenting to the emergency department (ED) of Hennepin County Medical Center in Minneapolis (HCMC Minneapolis, MN, USA), a tertiary care hospital receiving approximately 100,000 emergency patients per year. Approval was obtained from HCMC institutional review board. The HIPAA Methods for De-identification of Protected Health Information were used.

All ED ECGs were recorded on Mortara machines, with the incorporated Veritas® automated interpretation immediately performed. Thus, the Veritas® algorithm was the only contemporary algorithm with which we could compare the DNN analysis system of Cardiologs®.

In order for this study to be representative of ECG application in an ED, we randomly selected ECGs recorded in the ED of HCMC. The electronic medical record (Epic) data

management system at HCMC was searched for all ECGs recorded in the ED on patients aged 18 or older from July 1, 2015 to December 31, 2015. All HCMC ECGs used to train the DNN were from before this date; thus, none of the study ECGs had been used to train the DNN. From these ECGs, 1,500 were randomly selected. The data includes the ECG signal in the raw digital format, the automated Veritas® algorithm ECG interpretation from the Mortara® Instruments ECG machines (Mortara, Milwaukee, WI), the physician confirmation ("overread") interpretation, and the department in which the ECG was recorded (Emergency Department, etc.).

The algorithm from Cardiologs predicts diagnostics from raw 12-lead ECG electrode recordings. It outputs the probability of presence of 76 different labels. These labels can correspond both to general classes of pathologies (e.g. ventricular rhythm) or to specific pathologies within these classes (e.g. ventricular tachycardia). Therefore, labels are non-exclusive, e.g. an ECG can be predicted to correspond both to ventricular rhythm and ventricular tachycardia. The final diagnostic consists in the list of labels with probability superior to 0.5. Importantly, we use a single model to predict the presence or absence of all labels simultaneously. This enables the model to take into account the dependence between pathologies. For example, when a patient has a pre-excitation syndrome, the repolarization is affected by this pre-excitation. The algorithm can then learn to disregard the ST elevation of a pseudoinfarction pattern in order to avoid the false alarm of STEMI.

The algorithm consists in a convolutional neural network with 16 layers, with the first 13 being convolutional layers, followed by 3 fully connected layers.[15] This is similar to VGG, a neural network used for computer vision.[16] The model was implemented in TensorFlow.[17]

The model was trained using 100,000 ECG recordings for which the diagnostic was annotated by expert cardiologists. The training was done using stochastic gradient descent.[15] Briefly, at each training step, the model makes a prediction for an ECG, and this prediction is compared to the annotation from the expert cardiologist. Then, the parameters of the model are adjusted in order to lower the difference between predictions and annotations. This procedure is repeated for all ECGs in the training set, using each ECG multiple times. At testing time, the model is used to make predictions for each ECG in the testing set, and model predictions are compared against annotations. Importantly, the model was trained using data from multiple recording devices, and only a small portion of training examples were recorded using Mortara® ECG machine (5,000 out of 80,000). Therefore, our algorithm is expected to be relatively device independent, and not limited to Mortara® ECG machine.

By comparison, the Veritas algorithm is embedded on the Mortara's ECG machine. Briefly, the Veritas algorithm consists in two steps: beats are first detected and compared to a representative beat generated using all artifact-free complexes from all 12 leads in the 10-second ECG recording.[18] Then, arrhythmia detection as well as QRS-T-U analysis are performed by extracting landmarks with proprietary methods. Such methods have not been publicly disclosed, except for QT/RR interval measurements.[18-20]

Some previous studies have attempted to measure the performance of multi-label prediction in ECG diagnostics.[21,22] However, methodologies was restricted to a small number of possibly combined pathologies. For instance, studies of the accuracy of the Glasgow 12-lead ECG analysis program combined ventricular hypertrophy and myocardial infarction.[21] Willems et al. restricted experts' annotations to single labels.[22] Therefore, as stated above, no prior study has compared the performance of different algorithms in the global interpretation (both

rhythm and QRS-T-U analysis) of unselected 12-lead ECGs, and so there was no precedent for methodology. Previous studies were limited to a small set of labels, including myocardial infarction and hypertrophies, but not to global ECG intepretation. But there are over 100 possible discrete labels (e.g., left bundle branch block, LBBB) that can be applied to any ECG, if present. Different algorithms do not have the same label (diagnosis) or set of labels for each ECG abnormality, and this complicates any comparison. Thus, it was necessary to create a new model for comparison. To do so, we prospectively mapped all labels for each algorithm to a standard set of groups.

It is not straightforward to define sensitivity and specificity in the case of multiple labels. Different alternatives exist, depending on how we define the "correctness" of a prediction. One could define a prediction to be correct if all pathologies and sub-pathologies are detected, and precisely detected. Using this definition, a prediction of "Bundle Branch Block" would be judged incorrect in the case of Left Bundle Branch Block because the prediction is not specific enough. However, as shown by this example, most predictions would be judged incorrect if they are not perfectly precise, even if they contain relevant information (here, "Bundle Branch Block" is more informative than "No Bundle Branch Block"). Alternatively, one could measure the correctness of predictions for each pathology individually. However, the specificity for infrequent labels such as "Left Bundle Branch Block," would be always close to 100%, even for a naive algorithm always predicting "No Left Bundle Branch Block."

In order to manage these obstacles, we prospectively agreed on a list of 17 groups of cardiac abnormalities able to express the whole range of possible interpretations at a less granular

level than individual abnormalities (**Table 1**). This grouping made it possible to have a unified system for comparison of two different analysis systems. Furthermore, it helps to prevent ambiguities; for example, atrial fibrillation and flutter, which are often difficult for algorithms to differentiate, are grouped together (and then could be further classified into "significant" or "emergency" rhythms depending on a heart rate greater than, or less than, 120 beats per minute). STEMI and pericarditis, which are similarly difficult for algorithms to differentiate, are grouped together as acute emergencies. An automatic mapping from the two different annotation systems to this reduced list of groups was prospectively designed prior to the study initiation.

Furthermore, for the reasons outlined above, the statistical category of "true negative" for either individual labels or for the 17 groups was not applied; thus, we could not assess either specificity or negative predictive value. Analysis focused on the more meaningful categories of sensitivity and positive predictive value for identification of one or more of the 17 groups on each ECG.

We designed a comparison following the logical thought process of a reviewer in an emergency room where an ECG is deemed to be necessary. Is this ECG normal or abnormal? If it is abnormal, is it an emergency? If it is abnormal but not an emergency, is it significant or non-significant? Therefore, each of the 17 groups of conditions (16 abnormals, and 1 group for normal) (**Table 1**) was prospectively assigned an acuity category agreed upon by authors: "emergency" for abnormalities which require emergent treatment or consideration, "significant" for abnormalities that may have clinical significance to the emergency physician, but do not require immediate action, "non-significant" for abnormalities that are minor or non-specific, and "normal" for all normal or normal variant labels, including normal sinus rhythm

and normal variant ST Elevation (also often called "early repolarization"). "Emergency" and "significant" abnormalities were together called "major" abnormalities.

Each ECG could have features of one or more groups. The final classification of each ECG into any category was made according to the most serious classification. For instance, if any group was in the emergency classification, even if others were not, then that ECG would be classified as "Emergency." For a more specific example: atrial fibrillation at a rate of 130/min with a left bundle branch block is classified as "emergency" because of the "rhythm emergency" (atrial fibrillation with rate ≥ 120), in spite of the fact that "bundle branch block" is only classified as "significant." Similarly, if the most severe feature was "significant", the final classification was significant even if all other features were "non-significant" or "normal."

Since in acuity classifications (in contrast to the 16 abnormal groups) there was only one outcome for statistical analysis (major abnormality vs. none), acuity classifications could be assessed for true negatives and thus specificity and negative predictive value (NPV) could be calculated for this measure.

All ECGs were directly analyzed by the Veritas® algorithm at the time of ECG recording. The digital ECG signal was then fed to the Cardiologs® algorithm, which yielded an independent interpretation. For each ECG, the output report of each algorithm was mapped to the groups that had been chosen, as above.

The primary outcome was the performance of the algorithms in finding at least one "major" abnormality. The secondary outcome was the proportion of all ECGs for which all groups

were identified, with no false negative or false positive groups ("accurate ECG interpretation"). Additionally, we measured sensitivity and positive predictive value (PPV) for any abnormal group. Finally, we also assessed the performance in identifying emergency groups.

Reference standard

Two experienced, blinded interpreters independently assessed each of 1500 ECGs. The interpreters chose all abnormalities identified on each ECG from a pick list, whether of rhythm, QRS, or ST-T-U waves. To this end, all ECGs were displayed on an ECG Analysis platform designed for this study, allowing the annotators to:

1) Visualize the ECG, with access to information such as the patient age, estimated delineation (the computerized onsets and offsets of waves), and standard measurements [heart rate, P-wave duration, P-wave axis, PR interval, QRS duration and axis, QT interval, and corrected QT (Fridericia)].

2) Manipulate the ECG in various ways, such as displaying the full 10 seconds of each lead when desired, decreasing the amplitude of the signals, measuring amplitudes and duration, and zooming in to observe fine details.

3) Provide a rhythm and QRS-T-U analysis interpretation.

Each interpreter was trained to use this ECG Analysis Platform by videoconference, and was given the ECG Analysis Platform Instructions for Use.

If, and only if, the initial expert interpretation resulted in a classification into discrepant "groups," (See **Table 1** for the definitions of the 17 groups), the ECG was reviewed by a third ("tiebreaker") expert. If the tiebreaker disagreed with both interpretations, a consensus was reached with all interpreters participating. This final read was the reference standard, against which the automated interpretations were assessed. ECGs were excluded by interpreters if, in their opinion, the quality was inadequate or if there was lead misplacement. Importantly, none of these three interpreters had any access to Cardiologs' or Veritas' predicted diagnosis.

Statistics

The level of agreement comparing initial interpreters, and also comparing initial interpreters to the reference standard, was evaluated with percent agreement and kappa coefficient (with 95% CIs). Figures are given in proportions with 95% confidence intervals (CIs). Two-sided chi square tests were used to assess significance. McNemar's test was used to compare sensitivities.

Sample size was calculated based on analysis of 2.5 years of HCMC Emergency Department ECGs, in which the prevalence of major abnormalities (emergency findings plus significant findings) was 25%. Assuming a minimum value of 85% for either sensitivity or specificity, a sample of 1,500 patients would give an acceptable 95% confidence interval of 81% to 88% for the sensitivity.

Results

Of 24,123 ECGs recorded during the time period, 1,500 ECGs were randomly selected; 27 were excluded because of artifact or lead misplacement, leaving 1,473 ECGs. 559 interpretations (37.9%) resulted in discordant classification for one or more of the 17 groups

and thus needed tiebreaking; hence, agreement on the full set of groups of initial interpreters was 62.1% (95CI 60.3-63.8). Among discrepant ECGs, 239 resulted in discrepant classification into normal vs. non-significant, and 7 required a consensus discussion. Out of 3,191 initially annotated groups through the 2 interpretation rounds, 2,124 were matches between both interpreters (66.6%, 95CI 64.9-68.2). Kappa between initial interpreters on acuity categories was 0.55 (moderate agreement). Agreement on acuity categories was 68.8% (66.4-71.2). Kappa between reference annotations and the initial interpreters was from 0.72 (substantial) to 0.81 (near perfect), and agreement ranged from 81% to 87%. **Table 2** shows the count of abnormal groups.

Fifty-two (3.5%) of 1,473 ECGs had at least one emergent abnormality, for a total of 60 emergencies. There were 445 ECGs (30.2%) with 550 "significant" abnormalities. Thus, 497 ECGs (33.7%) had 610 "major abnormalities." There were 402 (27.3%) ECGs in the non-significant category, and 574 ECGs (39.0%) in the normal category. The largest of the 5 emergency groups was "other acute emergencies (STEMI, NSTEMI, pericarditis etc.)," with 33 of the total of 60 emergencies. Of the 445 ECGs with "significant" abnormalities, the most common by far was "significant sinus rhythm" due to sinus tachycardia (n = 259).

Primary Outcome: Performance of Classification of ECGs into Major Abnormality

See Table 3. Table 3 shows the algorithm performance of acuity classification for major abnormalities. Accuracy was 77.6% (95% CI: 75.4-79.7) for Cardiologs® and 68.3% (95% CI: 65.9-70.6) for Veritas® ($p < 0.0001$). Because Cardiologs® had approximately half as many false positives, accuracy and especially specificity (94.0% vs. 84.7%) and PPV (88.2% vs. 75.4%) were significantly higher for Cardiologs®.

For acuity classification, Veritas® trended towards better sensitivity (92.0% vs. 88.7%, (p = 0.086), with more true positives (457 vs. 441). However, in 9.2% (42/457) of those true positives, the group was incorrectly identified, compared with only 4.1% (18/441) for Cardiologs®. Thus, in approximately 9% of true positives for "major abnormality," vs. 4% for Cardiologs®, Veritas® made the correct classification of major abnormality through an incorrect diagnosis. This result is due to the fact that classifying an ECG as having an emergency abnormality did not necessarily imply that the found abnormality was correct: for instance, incorrectly finding ventricular tachycardia instead of STEMI would lead to a true positive. In fact, neither algorithm specifically identified the two ventricular rhythm emergencies, but both identified these ECGs overall as "emergency." Interestingly, both ECGs not only had initial discrepant interpretations but required consensus (See **Table 5).**

Secondary Outcome: Correct identification of Groups.

See Tables 4 a-c. Table 4a shows groups correctly identified on all ECGs, on ECGs with ≥ 1 abnormality, ECGs with ≥ 1 major abnormality, and ≥ 1 emergency; Cardiologs® significantly outperformed Veritas® on all measures. Table 4b shows sensitivity for all abnormalities, for major abnormalities, and for emergencies; the algorithms were not significantly different. Finally, Table 4c shows PPV for all abnormalities, for major abnormalities, and for emergencies; because it had many fewer false positives, Cardiologs® significantly outperformed Veritas® in all categories.

We manually examined the false negatives at the label level, in detail (See **Table 6a**). The largest contingent of false negatives for both algorithms were clearly due to Non-STEMI. We similarly examined false positive emergencies (**Table 6b**). There were 15 false positive

supraventricular rhythm emergencies for Veritas® vs. 5 for Cardiologs®. This reflects the difficulty for the Veritas® algorithm in differentiating between sinus tachycardia and other supraventricular rhythms. Additionally, Veritas®' measurements of QTc were often erroneous, and the algorithm overcalled STEMIs.

Congruent vs. discrepant ECGs

ECGs with initial discrepant interpretations (requiring tiebreaking) may be difficult and borderline ECGs; therefore, we compared the algorithms on the 914 congruent ECGs that had no discrepant groups (i.e., had concordant initial expert interpretations, and thus did *not* require tiebreaking). See Tables 7 and 8a-c. For this group, the results mirror the larger cohort, except that in this group, in addition to having higher specificity and PPV, Cardiologs® also was more sensitive for abnormal ECGs and also for abnormal ECGs with at least 1 major abnormality.

Discussion

Cardiologs has developed the first neural network able to detect multiple heart conditions simultaneously, using only the raw ECG signal as input. Here, we measured its performance for ECG applications representative of an emergency department. Compared to a reference standard, for classifying ED ECGs into major abnormality (emergency or significant), the Cardiologs®' DNN algorithm was significantly more accurate (92.2% vs. 87.2%), with significantly better specificity (94.0% vs. 84.7%) and PPV (88.2% vs. 75.4%) and a false discovery rate less than half of Veritas' (11.8% vs. 24.6%). Moreover, it was superior to the conventional algorithm in accurately identifying groups of abnormalities, with similar sensitivity (69.6% vs. 68.3%), and significantly better PPV (74.0% vs. 56.5%). Cardiologs identified the correct set of groups for 72.0% of ECGs, while Veritas identified only 59.8%.

For the subset of 914 ECGs that did not require tiebreaking, Cardiologs performed with higher sensitivity (84.1% vs. 78.8%), in addition to a higher PPV, for accurately identifying groups. Since Veritas is embedded in the Mortara machine, Veritas might be expected to perform better than Cardiologs; therefore, the better performance of Cardiologs, which is fully device-independent, is particularly noteworthy.

Over-diagnosis is recognized as a particular problem in computer-interpreted electrocardiograms.[2,7] Thus, although sensitivity for life threatening disorders is critical, specificity and PPV are also very important.

This study is the first to show the improved performance of a deep learning algorithm over a standard ECG interpretation algorithm on unselected 12-lead ECGs, although the improvement is modest. It may, in fact, be the first published direct comparison of automated ECG algorithms.[2] Interesting questions arise from this study, in particular concerning what is the ECG reference standard for some structural abnormalities (QRS-T-U), and what would be the performance of actual physicians, either overreading the algorithm or not, compared to the algorithms.

ECGs are interpreted in the clinical context by non-experts in real time with the aid of the automated interpretation. We know that the final ECG interpretation is greatly influenced by the automated interpretation, and sometimes in error, with clinical consequences.[1,5-8] An ideal study, in the clinical context, would be to randomize the automated interpretation to the DNN vs. contemporary algorithm. We could then use the final clinical physician over-read for comparison to an expert reference standard, or even to outcomes and patient management, to

assess carefully ruling in one possibility and ruling out the other, based on clinical and laboratory data.

Previous neural networks for ECG interpretation had many limitations. Most neural networks, though they can also learn on their own, are trained to find ECG abnormalities based on input of a simplified representation of an ECG, with handcrafted features such as the ST elevation value.[23,24] Others have a single simplified output such as acute MI,[23,24] or AV block.[25] Others have a list of exclusive outputs defining an ECG or a beat [such as left bundle branch block (LBBB) vs. right BBB (RBBB) vs. paced vs. ventricular vs. normal vs. other].[26-28] Others analyze rhythm only, and one exceeded the performance of board certified cardiologists for this task.[12]

Cardiologs®' algorithm is the first deep neural network algorithm that uses directly the whole 12-lead ECG as input, without any instructions, and interprets the whole 12-lead ECG. One additional reason for its success is that this neural network, unlike the conventional algorithms, has "non-exclusive" labels: this means that the single neural network algorithm can produce several labels (one for each abnormality) on a unique ECG, whereas the conventional algorithms use a separate sub-algorithm for each abnormality or set of disjointed abnormalities. For instance, if an ECG has both "atrial fibrillation" and "left bundle branch block," the quality of the interpretation is improved since the detection of a left bundle branch block can then help the algorithm to differentiate sinus rhythm from a ventricular rhythm. The conventional approach is to have two separate instructions (two different algorithms) for these two different outputs and to produce independent labels that are unrelated. In such a conventional algorithm, one label does not contribute to the diagnosis of another label, although simple rules can deactivate labels when the combination does not make sense (e.g.,

bundle branch blocks in case of ventricular tachycardia). It may diagnose both atrial fibrillation and left bundle branch block, but the diagnosis of either one does not affect the diagnosis of the other.

The neural network directly produces 75 "non-exclusive" labels (other labels are added afterwards based on measurements such as heart rate and QT intervals). Because of these non-exclusive labels, it can also produce outputs of different granularity. For ECGs with a high degree of difficulty, it was designed to produce a more general (less granular) interpretation. For instance, it may give the less precise interpretation of "atrial fibrillation or flutter" in a case of atrial fibrillation that is difficult to differentiate from atrial flutter. Handling these different levels of granularity was done using an ontology of cardiovascular diseases. Similar methods yielded positive results in other fields such as dermatology.[10]

Neural networks need a training phase in order to tune their numerous parameters so that it outputs an accurate interpretation when an ECG is provided as input. The interpretation is expressed as a sequence of values representing the probability of presence of each abnormality in the whole ECG. For instance, if a premature atrial complex (PAC) is present anywhere in the ECG, the sequence will contain a “1” at the index corresponding to the “PAC” label. Only the presence or absence of the abnormalities is given to the algorithm; no information is given on the location of the abnormality, or on the reason that it was interpreted as such. During each step of training, an ECG is provided to the neural network, which outputs an interpretation. This interpretation is then compared to the expected interpretation. The parameters of the neural network are slightly modified so that the interpretation gets closer to the expected interpretation. For this first version, this process was repeated countless times with approximately 130,000 ECGs and their corresponding interpretations. ECGs have been and will be continually added in the future to further refine the algorithm, so that it is

always improving. As of the end of 2017, the DNN has been trained on approximately 170,000 ECGs.

Limitations

The most important limitation was that the two algorithms do not have exactly identical individual labels. Thus, the labels from each had to be mapped to the 17 groups. Furthermore, many groups were represented by small numbers, including the absence of any ventricular rhythms and only 3 AV conduction disturbances. Emergencies had particularly low numbers (total, 60), which made it difficult to meaningfully compare and contrast the sensitivities of the algorithms for emergencies, even with such a large dataset. This limitation is the direct result of the main strength of the study: it represents a random sample of real-life ED ECGs rather than a selection of highly abnormal ECGs. Moreover, the large number of normal and non-significant ECGs made for a powerful comparison of specificity and PPV. Some groups of emergencies were heterogeneous, such as grouping STEMI with Pericarditis; since these may be difficult to differentiate for both clinicians and algorithms, we grouped them primarily to assess the algorithms' recognition of an emergency, with the rationale that once warned of the presence of emergency, the clinician would be prompted to assess carefully.

Since the experts used Cardiologs®' platform for the annotations, they may have also been biased by the measurements (of QT interval, or QRS duration, or heart rate) that were provided to them. Indeed, these measurements are the ones used by the algorithm for measurement-based abnormalities such as the PR interval. Experts were however also provided with tools for checking the measurements themselves in order to thwart this bias,

which is limited to only a few labels. More importantly, very few major abnormalities or accurate interpretations were dependent on these measurements.

Another limitation was the reference standard, which depended on expert interpretation. This limitation is particularly acute for Non-STEMI. We did not have anatomic or physiologic outcomes data for each ECG. However, in clinical practice, ECGs are interpreted by machines with human oversight. In future studies, the reference standard should perhaps always be consensus of the experts.

Because all ECGs in the training and testing set were recorded before July 1, 2015, and all used for this study were from after that date, no ECG that was used for training could have been used in the study. However, because of the need for ECG de-identification for patient privacy, we could not be certain that a patient whose ECG was in the study did not have a different previous ECG recorded that had been used for training. However, this is of little consequence for two reasons. First, there would have been little overlap because the testing set is composed of 1,500 ECGs, which were sampled randomly amongst 30,000 ECGs, and the training set was composed of 5,000 ECGs from Mortara's device, which were selected out of 80,000 possible ECGs. Second, ECGs from one individual are not perfectly constant over time; they change with changing patient condition, and would be unlikely to be substantially similar.

Finally, it would be interesting to compare the performance of Cardiologs' solution of ECG analysis in the ED to other existing algorithms. However, this was hindered by multiple factors. First, other solutions are either embedded on recording devices (e.g., GE Marquette™ 12SL or Philips DXL) or proprietary (like Glasgow® by Physio-Control).[21] In both cases, we

did not have access to these algorithms' predictions for our ECGs from an ED. Second, we could have used performance statistics reported in other previous studies. However, contrary to standard belief, sensitivity and specificity do differ in populations with widely differing prevalence of pathologies.[29] Here, in order to make this study as meaningful as possible in the context of emergency medicine, we studied a population representative of patients presenting to an ED, in which the prevalence of each entity is very low. In previous studies, the prevalence of each studied entity was either very high, or only one element (e.g, acute myocardial infarction) was studied.[21,23,30,31] Therefore, we expect to measure different performances than those reported in studies with different populations. Historical controls (i.e., performance characteristics from other studies) are furthermore not appropriate because no two ECGs are the same, which would render comparison inaccurate. Third, we propose a unique algorithm making predictions for all of multiple pathologies simultaneously. However, aside from Veritas, other methods omit many or even most pathologies that are predicted by Cardiologs' solution, which again makes comparison impossible.[32-40] While some pathologies like atrial fibrillation are often addressed, [32-34] some other pathologies such as idioventricular rhythm or pericarditis are rarely addressed. Therefore, most proposed methods are not suitable for comparison with Cardiologs' algorithm concerning simultaneous detection of multiple pathologies.

Conclusion

The Cardiologs®' ECG algorithm, the first deep neural network automated 12-lead ECG interpretation algorithm, performed significantly better than Veritas®' algorithm, identifying ECGs with major abnormalities with a higher positive predictive value while maintaining equal sensitivity. It had a significantly higher rate of accurate ECG interpretation, with similar sensitivity and higher PPV.

References

1. Hughes KE, Lewis SM, Katz L, Jones J. Safety of Computer Interpretation of Normal Triage Electrocardiograms. Acad Emerg Med 2017;24:120-124.
2. Schlapfer J, Wellens HJ. Computer-Interpreted Electrocardiograms: Benefits and Limitations. Journal of the American College of Cardiology 2017;70:1183-1192.
3. Mawri S, Michaels A, Gibbs J, Shah S, Rao S, Kugelmass A, Lingam N, Arida M, Jacobsen G, Rowlandson I, Iyer K, A. K, Mccord J. The Comparison of Physician to Computer Interpreted Electrocardiograms on ST-elevation Myocardial Infarction Door-to-balloon Times. Critical Pathways in Cardiology 2016;15:22-25.
4. Garvey JL, Zegre-Hemsey J, Gregg RE, Studnek JR. Electrocardiographic diagnosis of ST segment elevation myocardial infarction: An evaluation of three automated interpretation algorithms Journal of Electrocardiology 2016;49:728-732.
5. Novotny T, Bond R, Andrsova I, Koc L, Sisakova M, Finlay D, Guldenring D, Spinar J, Malik M. The role of computerized diagnostic proposals in the interpretation of the 12-lead electrocardiogram by cardiology and non-cardiology fellows. Int J Med Inform 2017;101:85-92.
6. Martinez-Losas P, Higueras J, Gomez-Polo JC, Brabyn P, Ferrer JM, Canadas V, Villacastin JP. The influence of computerized interpretation of an electrocardiogram reading. Am J Emerg Med 2016;34:2031-2032.
7. Bogun F, Anh D, Kalahasty G, Wissner E, Bou Serhal C, Bazzi R, Weaver WD, Schuger C. Misdiagnosis of atrial fibrillation and its clinical consequences. Am J Med 2004;117:636-642.
8. Anh D, Krishnan S, Bogun F. Accuracy of electrocardiogram interpretation by cardiologists in the setting of incorrect computer analysis. J Electrocardiol 2006;39:343-345.
9. Madias JE. Computerized interpretation of electrocardiograms: Taking stock and implementing new knowledge. J Electrocardiol 2018;51:413-415.
10. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks Nature Research 2017;542:115-118.
11. Eric Topol's Top 10 Tech Advances Shaping Medicine. Medscape, 2018. (Accessed January 16, 2018, at https://www.medscape.com/viewarticle/890982?nlid=120069_3869&src=WNL_mdplsfeat_180116_mscpedit_card&uac=212085PG&spon=2&impID=1534991&faf=1.)
12. Rajpurkar P, Hannun AY, Haghpanahi M, Bourn C, Ng AY. Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks. 2017.
13. Li J, Rapin J, roser A, Smith SW, Fleureau Y, Taboulet P. Deep neural networks improve atrial fibrillation detection in Holter: first results. European Journal of Preventive Cardiology 2016;23:41.
14. Smith SW, Rapin J, Li J, Walsh BM, Rosier A, Fiorina L, Dodd KW, Taboulet P. Improved Interpretation of Atrial Dysrhythmias by a New Neural Network

Electrocardiogram Interpretation Algorithm. Abstract 670. Acad Emerg Med 2017;24:S235.
15. Bengio Y, Courville A, Vincent P. Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence 2013;35:1798-1828.
16. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. International Conference on Learning Representations; 2014.
17. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mane D, Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker, Vanhoucke PV, Vasudevan V, Viegas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng X. Large-Scale Machine Learning on Heterogeneous Distributed Systems. Software available from tensorfloworg2015.
18. Mortara DW. Automated QT Measurement and Application to Detection of Moxifloxacin-Induced Changes. . Annals of Noninvasive Electrocardiology 2009;14:S30-S34.
19. Kligfield P, Badilini F, Denjoy I, Babaeizadeh S, Clark E, De Bie J, Devine B, Extramiana F, Generali G, Gregg R, Helfenbein E, Kors J, Leber R, Macfarlane P, Maison-Blanche P, Rowlandson I, Schmid R, Vaglio M, van Herpen G, Xue J, Young B, Green CL. Comparison of automated interval measurements by widely used algorithms in digital electrocardiographs. . Am Heart J 2018;200:1-10.
20. Borodin A, Pogorelov A, Zavyalova Y. Overview of algorithms for electrocardiograms analysis. . 13th Conference of Open Innovations Association (FRUCT) 2013 April 22-26; Petrozavodsk, Russia. p. 14-19.
21. Physio-Control I. Statement of Validation and Accuracy for the Glasgow® 12-Lead ECG Analysis Program version 27. 2009.
22. Willems JL, Abreu-Lima C, Arnaud P, et al. The diagnostic performance of computer programs for the interpretation of electrocardiograms. N Engl J Med 1991;325:1767-1773.
23. Heden B, Ohlin H, Rittner R, Edenbrandt L. Acute Myocardial Infarction Detected in the 12-Lead ECG by Artificial Neural Networks. Circulation 1997;96:1798-1802.
24. Kojuri J, Boostani R, Dehghani P, Nowroozipour F, Saki N. Prediction of acute myocardial infarction with artificial neural networks in patients with nondiagnostic electrocardiogram. Journal of Cardiovascular Disease Research 2015;6:51-59.
25. Meghriche S, Draa A, Boulemden M. On The Analysis of a Compound Neural Network for Detecting AtrioVentricular Heart Block (AVB) in an ECG Signal International Journal of Biological and Life Sciences 2008;4:1-11.
26. Bortolan G, Degani R, Willems JL. Neural networks for ECG classification Computers in Cardiology; 1990 23-26 September. p. 269-272.
27. Rai H, Trivedi A, Shukla S, Dubey V. ECG arrhythmia classification using daubechies wavelet and radial basis function neural network 2012 Nirma University International Conference on Engineering (NUiCONE); 2012. p. 1-6.
28. Kiranyz S, Ince T, Hamila R, Gabboui M. Convolutional Neural Networks for patient-specific ECG classification. 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society; 2015. p. 2608-2611.

29. Leeflang MMG, Rutjes AWS, Reitsma JB, Hooft L, Bossuyt P. Variation of a test's sensitivity and specificity with disease prevalence. Canadian Medical Association Journal 2013;185:E537-E544.
30. Gregg RE, Zhou SH, Dubin AM. Automated detection of ventricular pre-excitation in pediatric 12-lead ECG. J Electrocardiol 2016;49:37-41.
31. Kojuri J, Boostani R, Dehghani P, Nowroozipour F, Saki N. Prediction of acute myocardial infarction with artificial neural networks in patients with nondiagnostic electrocardiogram. Journal of Cardiovascular Disease Research 2015;6:51-59.
32. Kara S, Okandan M. Atrial fibrillation classification with artificial neural networks. . Pattern Recognition 2017;40:2967-2973.
33. Ten-Fan Yang T-F, Devine B, Macfarlane P. Artificial neural networks for the diagnosis of atrial fibrillation. . Medical and Biological Engineering and Computing 1994;32:615-619.
34. Xia Y, Wulan N, Wang K, Zhang H. Detecting atrial fibrillation by deep convolutional neural networks. Computers in Biology and Medicine 2018;93:84-92.
35. Zhang X-S, Zhu Y-S, Thakor NV, Wang Z-Z. Detecting ventricular tachycardia and fibrillation by complexity measure. . IEEE Transactions on Biomedical Engineering 1999;46:548-555.
36. Melillo P, Fusco R, Sansone M, Bracale M, Pecchia L. Discrimination power of long-term heart rate variability measures for chronic heart failure detection. Medical and Biological Engineering and Computing 2011;49:67-74.
37. Sayadi O, Shamsollahi MB. Life-threatening arrhythmia verification in ICU patients using the joint cardiovascular dynamical model and a Bayesian filter. IEEE Trans Biomed Eng 2011;58:2748-2757.
38. Tsipouras M, Fotiadis DI, Sideris D. An arrhythmia classification system based on the RR-interval signal. . Artif Intell Med 2005;33:237-250.
39. Kampouraki A, Nikou C, Manis G. Robustness of support vector machine-based classification of heart rate signals. . Conf Proc IEEE Eng Med Biol Soc; 2006. p. 2159–2162.
40. Jinkwon Kim, Hang Sik Shin, Kwangsoo Shin, Lee M. Robust algorithm for arrhythmia classification in ECG using extreme learning machine. . Biomed Eng Online 2009;8:1-12.

Table 1. Definitions of the groups and categories

Category	Group	Description
Emergent	Rhythm emergency	Sinoatrial block/sinus paralysis, non-sinus supraventricular and pacemaker rhythms with HR < 45 or HR > 120/min
Emergent	Ventricular rhythm emergency	Ventricular rhythm (including short runs) with HR < 45 or HR > 120/minute, indeterminate rhythms
Emergent	Atrioventricular (AV) conduction emergency	High-grade AV block: 2nd degree AV block Mobitz II or 3rd degree AV block
Emergent	Other acute emergency	Hyperkalemia, hypokalemia, acute ST Elevation MI, Acute Ischemia, recent MI, myocarditis/pericarditis
Emergent	Repolarization emergency	QTcF > 500 ms (QTc, Fridericia: QTc = QT/$\sqrt{3}$(RR) (QT divided by cube root of RR interval)
Significant	Significant sinus rhythm	Sinus rhythm with HR < 45 or HR > 100/minute
Significant	Significant rhythm	Non-sinus supraventricular and pacemaker rhythms with 45 < HR < 120/minute
Significant	Significant ventricular rhythm	Ventricular rhythm (including short runs) with 45 < HR < 120/minute
Significant	Significant AV conduction	Second degree AV block Mobitz I, Pre-excitation
Significant	Significant automatism	Supraventricular short runs, ventricular and atrial couplets/triplets/bigeminy/trigeminy
Significant	Significant bundle branch block	Complete right BBB, complete and incomplete left BBB, intraventricular conduction delay > 130 ms
Significant	Ventricular hypertrophy	Ventricular hypertrophies
Significant	Old myocardial infarction	Old (previous) myocardial infarction, any location
Significant	Repolarization suggesting myocardial ischemia	Repolarization suggesting myocardial ischemia
Significant	Other significant abnormalities	Rare pathologies (pulmonary diseases, Brugada, dextrocardia, etc.), QTcF > 470 and < 500 ms, abnormal U wave, short QTcF (< 320 ms)
Borderline	Non-significant	Isolated automatisms (premature atrial or ventricular beats), atrial hypertrophies, nonspecific abnormal QRS (left and right axes, fascicular blocks, IVCD > 110 ms,

		incomplete RBBB), non-specific repolarization abnormalities, borderline QTcF (430-470ms), first degree AV block, low voltage This label is used on abnormal ECGs in absence of any other label (if a major label is present, this label is omitted)
Normal	Normal	Sinus rhythm with heart rate between 45/minute and 100/minute, including morphological normal variants (early repolarization etc.).

Abbreviations for Table 1.
HR = heart rate
AV = atrio-ventricular
BBB = bundle branch block

Table 2. Count of abnormal groups

Group	Number	(%)
Rhythm emergency	19	1.3
Ventricular rhythm emergency	2	0.1
Other acute emergency	33	2.2
AV conduction emergency	2	0.1
Repolarization emergency	4	0.3
Significant sinus rhythm	259	17.6
Significant rhythm	47	3.2
Significant ventricular rhythm	0	0.0
Significant AV conduction	3	0.2
Significant automatism	20	1.4
Significant bundle branch block	66	4.5
Ventricular hypertrophy	32	2.2
Old myocardial infarction	92	6.2
Repolarization suggesting myocardial ischemia	13	0.9
Other significant abnormality	18	1.2
Nonsignificant	402	27.3
Normal	574	39.0

AV = atrio-ventricular

Note: because one ECG can have several groups of abnormalities, the figures of this table add up to more than 100%.

Table 3. Diagnostic Performance of Cardiologs® vs. Veritas® for identifying an ECG with at least one major abnormality (emergency or significant) vs. none (non-significant or Normal)

Major Abnormalities	**Cardiologs®, % (95% CI)**	**Veritas®, % (95% CI)**	**p-value**
Sensitivity	88.7 (85.7-91.2)	92.0 (89.2-94.0)	0.086
Specificity	94.0 (92.3-95.3)	84.7 (82.3-86.9)	<0.0001
PPV	88.2 (85.1-90.7)	75.4 (71.8-78.7)	<0.0001
NPV	94.2 (92.6-95.5)	95.4 (93.8-96.6)	0.27
Accuracy	92.2 (90.7-93.5)	87.2 (85.4-88.8)	<0.0001

PPV: positive predictive value; NPV: negative predictive value

Table 4a. ECGs with all Groups correctly identified (no overcalls and no undercalls)

	Cardiologs®		Veritas®		
	Number	%, 95% CI	Number	%, 95% CI	P
All ECGs	1060/1473	**72.0%** (69.6-74.2)	881/1473	**59.8%** (57.3-62.3)	< 0.0001
ECGs with ≥ 1 abnormality	567/899	**63.1%** (59.9-66.2)	517/899	**57.5%** (54.3-60.7)	0.016
ECGs with ≥ 1 Major abnormality	327/497	**65.8%** (61.5-69.8)	284/497	**57.1%** (52.8-61.4)	0.005
ECGs with ≥ 1 Emergency	24/52	**46.2%** (33.3-59.5)	23/52	**44.2%** (31.6-57.7)	NS

Table 4b. Sensitivity for Abnormal Groups

	Cardiologs®		Veritas®		
	Number	%, 95% CI	Number	%, 95% CI	P
All abnormalities	704/1012	**69.6%** (66.7-72.3)	691/1012	**68.3%** (65.3-71.1)	NS
Major abnormalities	464/610	**76.1%** (72.5-79.3)	458/610	**75.1%** (71.5-78.4)	NS
Emergency abnormalities	32/60	**53.3%** (40.9-65.4)	30/60	**50%** (37.7-62.3)	NS

Table 4c. Positive Predictive Value (True positives by the algorithm divided by all positives for that algorithm)

	Cardiologs®		Veritas®		
	Number	%, 95% CI	Number	%, 95% CI	P
All abnormalities	704/951	**74.0%** (71.1-76.7)	691/1223	**56.5%** (53.7-59.3)	< 0.0001
Major abnormalities	464/611	**75.9%** (72.4-79.2)	458/812	**56.4%** (53.0-59.8)	< 0.0001
Emergency abnormalities	32/52	**61.5%** (48.0-73.5)	30/73	**41.1%** (30.5-52.6)	0.024

Table 5. Manual investigation on the interpretation of the two ventricular emergency cases

	ECG1	ECG2
Cardiologs®	- Atrial flutter - 2nd or 3rd degree AVB - Complete right bundle branch block - QTc > 500ms	- Ischemia of indeterminate age - Atrial fibrillation - Ventricular pacemaker - QTc > 500 ms
Veritas®	- Complete right bundle branch block - Atrial tachycardia or flutter - Ischemia of indeterminate age - QTc > 470 ms	- Ventricular pacemaker - Acute Ischemia
Annotator 1	- Atrial flutter - Complete right bundle branch block	- Pacemaker - Non-sustained ventricular tachycardia
Annotator 2	- Atrial flutter - Complete right bundle branch block - QTc > 470 ms	- Ventricular pacemaker - QTc > 500 ms
Reference (through consensus)	- Atrial flutter - 3rd degree AV block - Ventricular escape rhythm - QTc > 470 ms	- Intraventricular conduction delay > 130ms - Acute Ischemia - Ventricular pacemaker - QTc > 500 ms - Non-sustained ventricular tachycardia

Table 6a. Manual investigation on the missed Emergencies

	Total	Cardiologs® (missed)	Veritas® (missed)
Supraventricular rhythm > 120 or < 45	19	1	0
Ventricular rhythm > 120 or < 45	2	0	0
Complete Heart Block	2	0	0
Hyperkalemia	2	0	1
Hypokalemia	1	1	1
STEMI	7	2	2
Non STEMI	20	16	13
Recent or Acute MI	2	1	1
Myocarditis/Pericarditis	1	0	0
QTcF > 500 ms	4	1	3
Total	60	22	21

Table 6b. Manual investigation on the false positives of Emergencies

	Cardiologs®	Veritas®
Supraventricular rhythm < 45 or > 120	5	15
2nd degree AV Block, Mobitz II	0	1
Hyperkalemia	1	1
STEMI	0	8
Non-STEMI	0	1
Recent MI	2	0
Pericarditis or myocarditis	5	5
QTc > 500 ms	5	11
Total	18	42

Table 7. Diagnostic Performance of Cardiologs® vs. Veritas® for identifying an ECG with at least one major abnormality, among the 914 cases that did not require tiebreaking.

Major Abnormalities	**Cardiologs®, % (95% CI)**	**Veritas®, % (95% CI)**	**p-value**
Sensitivity	98.0 (95.6-99.1)	96.6 (93.8-98.1)	0.22
Specificity	96.6 (94.9-97.8)	88.6 (85.8-90.8)	<0.0001
PPV	93.2 (89.8-95.5)	79.9 (75.5-83.8)	<0.0001
NPV	99.0 (97.9-99.5)	98.2 (96.7-99.0)	0.24
Accuracy	97.0 (95.7-98.0)	91.1 (89.1-92.8)	<0.0001

PPV: positive predictive value; NPV: negative predictive value

Table 8. Comparison of algorithms in Cases in which the initial expert interpretations were congruent (did not need tie-breaking). N = 914/1473 total (62.1%)

Table 8a. ECGs with all Groups correctly identified (no overcalls and no undercalls)

	Cardiologs®		Veritas®		
	Number	%, 95% CI	Number	%, 95% CI	P
All ECGs	779/914	85.2% (82.8-87.4)	645/914	70.6% (67.5-73.4)	< 0.0001
ECGs with ≥ 1 abnormality	400/504	79.4% (75.6-82.7)	353/504	70.0% (65.9-73.9)	0.0007
ECGs with ≥ 1 Major abnormality	249/293	85.0% (80.4-88.6)	216/293	73.7% (68.4-78.4)	0.0008
ECGs with ≥ 1 Emergency	17/21	81.0% (60.0-92.3)	15/21	71.4% (50.0-86.2)	0.4687

Table 8b. Sensitivity for Abnormal Groups

	Cardiologs®		Veritas®		
	Number	%, 95% CI	Number	%, 95% CI	P
All abnormalities	448/533	84.1% (80.7-86.9)	420/533	78.8% (75.1-82.1)	0.0001
Major abnormalities	297/322	92.2% (88.8-94.7)	283/322	87.9% (83.9-91.0)	0.0043
Emergency abnormalities	20/23	87.0% (67.9-95.5)	17/23	73.9% (53.5-87.5)	0.375

Table 8c. Positive Predictive Value (True positives by the algorithm divided by all positives for that algorithm)

	Cardiologs®		Veritas®		P
	Number	%, 95% CI	Number	%, 95% CI	P
All abnormalities	448/527	85.0% (81.7-87.8)	420/671	62.6% (58.9-66.2)	< 0.0001
Major abnormalities	297/347	85.6% (81.5-88.9)	283/439	64.5% (59.9-68.8)	< 0.0001
Emergency abnormalities	20/28	71.4% (52.9-84.7)	17/33	51.5% (35.2-67.5)	0.1126

Highlights

- We studied the first version of Cardiologs'® new deep neural network (DNN) 12-lead ECG interpretation algorithm, compared to a conventional algorithm
- Cardiologs' DNN was more accurate and specific in identifying ECGs with at least one major abnormality.
- Cardiologs' DNN had a significantly higher rate of accurate ECG interpretation, with higher specificity and PPV.
- Among non-controversial ECGs, the DNN also had higher sensitivity for abnormalities, in addition to higher specificity and accuracy.