Software-based detection of atrial fibrillation in long-term ECGs

Andreas Haeberlin, MD, DAS,† Laurent Roten, MD,* Manuel Schilling, MD,* Flavio Scarcia, MD,* Thomas Niederhauser, MSc,†† Rolf Vogel, MD, PhD.§ Juerg Fuhrer, MD,* Hildegard Tanner, MD†

From the †Department of Cardiology, Bern University Hospital and University of Bern, Bern, Switzerland, ‡ARTORG Center for Biomedical Engineering Research, University of Bern, Bern, Switzerland, §Institute for Human Centered Engineering – microLab, Engineering and Information Technology, Bern University of Applied Sciences, Biel, Switzerland and ‡Department of Cardiology, Bürgerspital Solothurn, Solothurn, Switzerland.

BACKGROUND Atrial fibrillation (AF) is common and may have severe consequences. Continuous long-term electrocardiogram (ECG) is widely used for AF screening. Recently, commercial ECG analysis software was launched, which automatically detects AF in long-term ECGs. It has been claimed that such tools offer reliable AF screening and save time for ECG analysis. However, this has not been investigated in a real-life patient cohort.

OBJECTIVE To investigate the performance of automatic software-based screening for AF in long-term ECGs.

METHODS Two independent physicians manually screened 22,601 hours of continuous long-term ECGs from 150 patients for AF. Presence, number, and duration of AF episodes were registered. Subsequently, the recordings were screened for AF by an established ECG analysis software (Pathfinder SL), and its performance was validated against the thorough manual analysis (gold standard).

RESULTS Sensitivity and specificity for AF detection was 98.5% (95% confidence interval 91.72%–99.96%) and 80.21% (95% confidence interval 70.83%–87.64%), respectively. Software-based AF detection was inferior to manual analysis by physicians (P < .0001). Median AF duration was underestimated (19.4 hours vs 22.1 hours; P < .001) and median number of AF episodes was overestimated (32 episodes vs 2 episodes; P < .001) by the software. In comparison to extensive quantitative manual ECG analysis, software-based analysis saved time (2 minutes vs 19 minutes; P < .001).

CONCLUSION Owing to its high sensitivity and ability to save time, software-based ECG analysis may be used as a screening tool for AF. An additional manual confirmatory analysis may be required to reduce the number of false-positive findings.

KEYWORDS Atrial fibrillation; Software; Software-based analysis; Atrial fibrillation detection; Long-term ECG

ABBREVIATIONS AF = atrial fibrillation; CI = confidence interval; ECG = electrocardiogram; IQR = interquartile range

(Heart Rhythm 2014;11:933–938) © 2014 Heart Rhythm Society. All rights reserved.

Introduction

Atrial fibrillation (AF) is common and associated with morbidity and increased mortality.1,2 However, therapeutic measures may prevent negative consequences, for example, oral anticoagulation lowers the risk of thromboembolic events. Thus, screening for AF is crucial to diagnose AF before the occurrence of adverse events. Continuous 7-day Holter electrocardiogram (ECG) is widely used and considered to be an effective method for AF detection.3 However, an analysis of such recordings can be time-consuming.4 To overcome this limitation, semi-automatic software-based analysis of continuous long-term ECG has been proposed for AF detection.5–7 Today, different manufacturers provide software tools, aiming to facilitate AF detection and save time for ECG analysis, which is of particular interest in the era of increasing AF prevalence and health-care expenditures. The implemented algorithms seem to distinguish AF well from sinus rhythm when evaluated using small selected ECG test databases.8–10 However, the task to reliably detect AF in continuous long-term ECGs may be substantially more challenging. A wide spectrum of arrhythmias has to be distinguished, and artifacts may perturb the analysis. In addition, fast data processing, visualization, and reliable analysis are crucial. The performance of an established state-of-the-art software to detect AF has not been investigated in a real-life patient cohort yet, and this was the purpose of the present study.

Methods

Study design

In this investigator-initiated single-center study, we validated the performance of a semiautomatic AF detection software...
against the extensive manual ECG analysis by physicians (gold standard) in 7-day ECG recordings. For this purpose, we retrospectively analyzed a data set of 161 continuous 7-day ECG recordings from 150 patients, included between November 2007 and January 2011. These recordings were acquired within 2 studies investigating (1) the performance of 2 different long-term ECG recorders or (2) the presence of AF in patients after aortic valve replacement.\(^4\,11\) The local ethics committee approved these studies (according to the Declaration of Helsinki), and all patients gave written informed consent.

**Study population**

All recordings were acquired from patients with known or suspected AF. Patient had to meet one of the following inclusion criteria:

1. Known AF, assessment before radiofrequency catheter ablation of AF
2. Assessment of AF recurrence 3, 6, or 12 months after radiofrequency catheter ablation
3. Suspected AF with symptoms highly suggestive of AF (history of irregular palpitations)
4. Screening for AF and atrioventricular conduction abnormalities after transcatheter or surgical aortic valve replacement

The baseline characteristics of the patients are summarized in Table 1.

**ECG recorder**

AF screening was performed using a long-term ECG recorder (Lifecard CF, Spacelabs Healthcare, Snoqualmie, WA, USA). This device continuously records 2 channels for 7 days with a sampling frequency of 1024 Hz. To register the 2 channels, 3 long-term surface electrodes were attached to each patient (BlueSensor VL, Ambu, Denmark). Both channels measured a bipolar lead. The first channel was derived between an electrode at the lower right border of the rib cage on the midclavicular line and an electrode at the lower left border of the rib cage on the anterior axillary line. The second channel was derived between an electrode at the lower right border of the rib cage on the midclavicular line and an electrode at the lower left border of the rib cage on the anterior axillary line.

**Manual ECG analysis**

Before automatic software-based ECG analysis (see the next paragraph), long-term ECGs were analyzed manually to reveal all registered episodes of AF. Specifically, all recordings were visualized by the Lifescreen software (Spacelabs Healthcare) to allow reviewing the original ECG traces for the presence of AF. First, this was done by careful direct visual analysis of the ECG traces minute after minute. In a second step, the recordings were reviewed again by analysis of R-R intervals vs time (the latter is called RR tachogram by the Lifescreen software) since R-R intervals show a high variability during AF and may thus help to identify AF episodes. The ECG analysis was performed by 2 independent physicians (1 senior electrophysiologist). In case of interrater disagreement, another senior electrophysiologist was asked to review the ECG and a decision was made by consensus. The goal of this extensive and thorough “manual analysis” was to establish a gold standard for the subsequent validation of software-based AF screening (see next paragraph). Quantitatively, the number and total duration of AF episodes were registered. AF was defined to last for ≥30 seconds and according to established standard definitions.\(^12\)

**Software-based AF screening**

Long-term ECGs were analyzed using the Pathfinder SL software (version 1.6.2, Spacelabs Healthcare). Each recording was automatically screened for AF episodes. The software was run without changing the AF detection default parameters (not possible).

Descriptively, the AF detection algorithm of the Pathfinder SL software works as follows (based on a technical document provided by the software retailer):

1. R-R intervals are detected.
2. R-R intervals <180 or >2000 ms are excluded, that is, considered as artifacts.
3. The ECG is split into slices of 20 seconds. The slices are filtered, and an R-R interval variability measure is calculated for each slice.
4. The variability is compared with (internal) thresholds. If it is higher than the threshold, an AF event is detected.
5. Nearby AF events are merged to avoid splitting up continuous AF episodes into many short single events.

The number of “AF episodes” (according to the software) and their total duration were registered. In a second step, all “AF episodes” were reviewed to judge the software’s classification (eg, true-positive or false-positive detection of AF). The overall time needed to run the analysis and review the episodes was registered.

---

**Table 1**

<table>
<thead>
<tr>
<th>Patient baseline characteristics</th>
<th>150</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (y)</td>
<td>70.8 ± 12.5</td>
</tr>
<tr>
<td>Sex: male</td>
<td>88 (58.7)</td>
</tr>
<tr>
<td>BMI (kg/m(^2))</td>
<td>26.1 ± 4.5</td>
</tr>
<tr>
<td>Arterial hypertension</td>
<td>95 (63.3)</td>
</tr>
<tr>
<td>Diabetes mellitus</td>
<td>23 (15.3)</td>
</tr>
<tr>
<td>CHADS(_2) score (median)</td>
<td>1.76 (2)</td>
</tr>
<tr>
<td>Ischemic heart disease (%)</td>
<td>58 (38.7)</td>
</tr>
<tr>
<td>Hypertensive heart disease (%)</td>
<td>35 (23.3)</td>
</tr>
<tr>
<td>Valvular heart disease</td>
<td>85 (57.0)</td>
</tr>
<tr>
<td>Left ventricular ejection fraction (%)</td>
<td>56.6 ± 11.3</td>
</tr>
<tr>
<td>Left atrial diameter, parasternal (mm)</td>
<td>44.2 ± 6.9</td>
</tr>
</tbody>
</table>

Values are presented as mean ± SD or as n (%).

BMI = body mass index.
Statistical analysis

For statistical analysis, SPSS version 21.0 (IBM, USA) and R version 3.0.1 for Windows were used. Categorical variables are expressed as numbers and percentages and continuous variables as mean ± SD or as median and interquartile range (IQR). 95% confidence intervals (CI) for sensitivity and specificity were calculated using the Clopper-Pearson method. The McNemar test was used to compare the performance of the software vs manual analysis in a 2 × 2 contingency table. Continuous variables were compared using a Wilcoxon signed-rank test. Correlation analysis was made using Spearman’s rank correlation coefficient (ρ_{Spearman}). A P value of ≤ .05 was considered to be significant.

Results

Registered AF episodes

We analyzed 161 continuous 7-day ECGs, with a median recording duration of 162.0 hours (IQR 140.5–166.8 hours; total duration 22,601 hours, ie, 2.6 years). AF was registered in 65 recordings (40.4%). The median duration of AF in these recordings was 22.1 hours (IQR 6.8–70.5 hours). Minimal AF duration was 32 seconds, and maximal duration was 169 hours. The median number of AF episodes per patient with AF was 2 (IQR 1–11.8 episodes; maximum 428 episodes).

Table 2  Contingency table of software-based AF detection

<table>
<thead>
<tr>
<th></th>
<th>AF diagnosed by manual analysis</th>
<th>No AF diagnosed by manual analysis</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>AF diagnosed by software</td>
<td>64</td>
<td>19</td>
<td>83</td>
</tr>
<tr>
<td>No AF diagnosed by software</td>
<td>1</td>
<td>77</td>
<td>78</td>
</tr>
<tr>
<td>Total</td>
<td>65</td>
<td>96</td>
<td>161</td>
</tr>
</tbody>
</table>

AF = atrial fibrillation.

Software-based detection of AF

Table 2 is a contingency table showing the software’s performance to detect AF. The software missed a single AF episode of 32 seconds (Figure 1). However, 19 patients without AF in the ECG registration were diagnosed with AF. Thus, the sensitivity for AF detection was 98.46% (95% CI 91.72%–99.96%) and specificity was 80.21% (95% CI 70.83%–87.64%). The positive predictive value was 77.1%, and the negative predictive value was 98.7%.

Some false-positive AF episodes (examples given in Figure 2) were episodes of sustained atrial tachycardia or atrial flutter, and thus clinically relevant (Figure 3). If those episodes would have been treated as true positives, this would have resulted in a sensitivity of 98.53% (95% CI 92.08%–99.96%) and a specificity of 82.80% (95% CI 73.57%–87.64%).

Comparison of software-based and manual detection of AF

Software-based detection of AF was inferior to manual analysis by physicians (P < .0001; Table 2). Prior manual analysis had not missed any AF episode diagnosed by the software.

In patients with AF, the software underestimated median AF duration (19.4 hours; IQR 3.3–45.8 hours) as compared with the true duration (22.1 hours; IQR 6.8–70.5 hours; P < .001). The correlation of the AF duration according to the software and to the manual analysis was ρ_{Spearman} = .79 (Figure 4, top panel). In contrast, the software overestimated the median number of AF episodes per patient with AF (32 episodes vs 2 episodes; IQR 10.5–95.5; P < .001). The correlation of the number of AF episodes according to the software and to the manual analysis was ρ_{Spearman} = .27 (Figure 4, bottom panel).

However, quantitative manual analysis required a median of 19 minutes (IQR 13–32 minutes) vs 2 minutes

Figure 1  Electrocardiogram of the 32-second lasting AF episode missed by the software (1 channel is shown at 25 mm/s). The beat annotation according to the software is shown: N = nonpremature beat; S = supraventricular premature beat; V = ventricular premature beat.
(IQR 2–5 minutes) to run the software and review the detected AF episodes (P < .001).

Discussion
We present for the first time results of a large-scale study investigating the performance of software-based screening for AF in continuous 7-day ECG recordings. A state-of-the-art software shows an inferior performance if validated against extensive manual ECG analysis (gold standard) but may save time for AF screening.

Performance of software-based screening for AF
The diagnosis of AF was missed in 1 recording, which showed a single 32-second AF episode (Figure 1). At the beginning of this episode, the beats were misclassified as ventricular and supraventricular premature beats (first and first half of the second line of Figure 1). The remaining part of the episode was not highly irregular and relatively short. Thus, the N-N variability may not have been higher than the internal “AF threshold” of the software (step 4 of the detection algorithm). This explains why this AF episode was missed.

Owing to the potentially devastating consequences of AF, it is crucial not to miss AF episodes. As must be expected from an automated tool, sensitivity was high, but not 100%. Thus, AF may be missed if the ECG is not reviewed manually. In contrast to sensitivity, we observed a limited specificity for AF episodes. However, a certain number of false-positive AF detections may be less severe than missing AF. In addition, 15.6% of the misclassified events were sustained atrial tachycardias or atrial flutter (Figure 3) and, thus, clinically relevant. Nevertheless, a careful review of all “AF episodes” is crucial to avoid AF overdiagnosis. In comparison to numerous papers reporting high sensitivity/specificity for different AF detecting algorithms,8–10 we observed a lower performance of the software. This can be explained by 2 reasons. First, such algorithms are usually validated by using test ECG databases such as the MIT-BIH arrhythmia database, consisting of just a few selected short recordings.14 This may have led to a certain “algorithmic overfitting” in other papers,8–10 that is, the algorithms may perform well in identifying these specific AF episodes but may fail to detect AF episodes from another population. Second, the specificity may be impaired owing to other arrhythmias such as frequent atrial premature beats or atrial tachycardias, which quite often are also present in patients with AF.15

Figure 2  Examples of 3 episodes (A-V sequential rhythm) that were misclassified as AF episodes: atrial premature beats with varying coupling intervals (top panel), nonsustained atrial tachycardia and atrial premature beat (middle panel), and combined QRS undersensing and T-wave oversensing (bottom panel). The beat annotation according to the software is shown: N = nonpremature beat; V = ventricular premature beat; the triangle in the top and middle panel annotates the beginning of the “AF episode.”

Figure 3  Bar plot summarizing the correct rhythm of episodes mistakenly classified as AF by the software. AF = atrial fibrillation; APB = atrial premature beat; artifact = QRS undersensing due to signal disturbances (6 episodes), pacemaker malfunction (1 episode), or combined T-wave oversensing/QRS undersensing (1 episode); AT/AFLU = atrial tachycardia and atrial flutter; NSAT = nonsustained atrial tachycardia; NSVT = nonsustained ventricular tachycardia; VPB = ventricular premature beat.
In addition to the inferior performance of the software regarding the diagnosis of AF as compared with manual analysis, the software showed limitations in quantifying the AF burden. We often found continuous episodes of AF split up by the software into several AF episodes and interrupted by sinus rhythm (eg, Figure 4, bottom panel, point at the top left corner). This explains why the software underestimates the total AF duration and overestimates the number of AF episodes in recordings with AF. However, the software mainly underestimates the duration of longer AF episodes (Figure 4, top panel). To the author’s knowledge, there are no data suggesting that AF episodes of, for example, 8000 minutes are less risky than AF episodes of, for example, 10,000 minutes. Thus, for clinical decision making, the AF duration underestimation by the software may be less relevant.

Predictive power of software-based AF screening in long-term ECGs

In our study cohort, we observed a negative predictive value of 98.7% and a positive predictive value of 77.1%. Since many of our patients had known AF, there was a high a priori chance to register many long-term ECGs with AF episodes. However, the prevalence of AF in patient cohorts typically screened for AF (eg, after ischemic stroke) may be lower. Thus, in a given population with a lower pretest probability of AF (assumed identical sensitivity and specificity), a higher negative predictive value but even lower positive predictive value would be observed.

Methodological limitations of software-based long-term ECG analysis

Software-based screening for AF in long-term ECGs suffers from a major technological limitation. Today’s detection algorithms analyze R-R intervals and their variability to evaluate ECGs for the presence of AF. QRS-peak detectors provide a reliable and robust detection of R peaks.16 In the surface ECG, the QRS complex differs from p and T waves with respect to not only amplitude but also the frequency components.17,18 Thus, stepwise band-pass filters are applied, amplifying R peaks and attenuating p- and T-waves by suppressing their frequency components.16,19 In contrast, implantable devices (eg, (S-)ICDs) require a more sophisticated approach. Detecting just “R peaks” is not sufficient for sustained ventricular tachycardia/ventricular tachycardia discrimination. The signal morphology has to be preserved to acquire signal templates, allowing good sustained ventricular tachycardia/ventricular tachycardia discrimination. However, this may also lead to a higher susceptibility to T-wave oversensing,16,20 which was a minor issue in our study (Figure 3, bottom panel).

As a key limitation of contemporary surface ECG peak detectors, the atrial activity is not detected or analyzed directly and independently from the QRS complex owing to the low amplitude of the p wave and a poor signal-to-noise ratio.8,9 Instead, R-R intervals are used as surrogate markers for the supraventricular activity and to detect atrial arrhythmias such as AF. This explains the limited specificity of the software: episodes that were misclassified as AF had another supraventricular origin in almost 95% of the cases (Figure 3). Since the atrial activity was not analyzed directly, the software algorithm failed. Particularly, frequent atrial premature beats with varying coupling intervals caused false-positive AF detection (Figures 2 and 3). If true premature complexes are not “premature enough” for the software, they are classified as “nonpremature beats” (Figure 2, top and middle panels). Such “normal” beats subsequently increase the variability measure (step 4 of the AF detection algorithm), and the software recognizes an “AF episode.”

Alternatively, esophageal long-term electrocardiography may provide a way out, since it records excellent atrial electrograms. This may allow reliable detection of the atrial activity, for example, in the case of atrial premature beats,21 and facilitate the diagnosis of AF.22
**ECG analysis duration**

Median duration of ECG analysis was more than 15 minutes shorter when using the software instead of performing a manual analysis. The rather long manual analysis time results from the quantitative manual analysis; that is, episodes had to be identified and counted and the duration of every episode was registered manually. Such an extensive analysis (used for validation purposes in this study) may be dispensable in clinical practice; a qualitative screening for AF may be sufficient, and thus time for manual analysis might be saved, although we hypothesize that software-based screening may be more consistent over time compared to human manual analysis. Because of mental fatigue (eg, after analyzing ECGs for a long time period), the analysis time of humans may increase and the reliability may decrease. However, this was not evaluated in the present study. Nevertheless, leaving out the manual analysis of the ECG trace and just running the software and reviewing the “AF episodes” may lead to an increase in efficiency.

**Study limitations**

We investigated the performance of 1 particular established state-of-the-art ECG analysis software (Pathfinder SL). Other AF detection algorithms may show a different performance. However, today’s software relies on the same methodological principle (RR analysis). Thus, a similar performance may be expected for other products. As a limitation of the software we used, AF detection parameters cannot be changed. Therefore, it is not possible to optimize the software default settings to achieve a higher performance.

We cannot rule out that some AF episodes have been missed, although we performed a careful manual review for AF episodes and no additional AF episodes were revealed by the software.

In addition, manual AF screening in real life outside a clinical study is often performed by technicians and not physicians (as done in this study). Furthermore, it may be less extensive in comparison to the detailed analysis we performed in this study for validation purposes. Thus, the performance of a software-based analysis compared to a real-life ECG analysis may be different (ie, likely be even better).

In 50 inpatients, the long-term ECG was acquired postoperatively, that is, the patients were immobilized, which may have reduced the number of motion artifacts that could have been misclassified as AF.

**Conclusion**

Software-based ECG analysis using the Pathfinder SL software shows a high sensitivity but limited specificity for revealing AF episodes. In addition, it may save time compared to manual analysis by humans. Thus, although not perfect, automated algorithms may be used as fairly reliable screening tools for AF. Manual analysis should be performed as a confirmatory analysis to reduce the number of false-positive findings.

**References**