Inter-System Agreement on P-Wave Measures Across Four Automated ECG Interpretation Algorithms

Meri Gjika[1], Agnese Sbrollini[1], Anders Björkelund[2], Claus Graff[3], Dillon J Dzikowicz[4], Fredrik Holmqvist[2], Junmo An[4], Peter W Macfarlane[5], Ramun Schmid[6], Laura Burattini[1], Pyotr G Platonov[2,*]

[1]Università Politecnica delle Marche, Ancona, Italy
[2]Lund University, Lund, Sweden
[3]Department of Health Science and Technology, Aalborg University, Aalborg, Denmark
[4]Philips, Cambridge, Massachusetts, USA
[5]University of Glasgow, Scotland, UK
[6]Schiller AG, Baar, Switzerland

Abstract

P-wave measures from the 12-lead ECG are clinical markers of atrial structure and remodeling in atrial cardiomyopathies. Automated ECG interpretation algorithms differ in P-wave quantification, and inter-system agreement is not fully characterized across commercial platforms. This work quantified inter-system agreement in P-wave measures across four algorithms: GE Marquette 12SL, Glasgow, Schiller ETM, and Philips DXL. The dataset comprised 140 anonymized 12-lead ECGs with P-wave morphological variability. The four algorithms independently processed the ECGs. Agreement was assessed using the intraclass correlation coefficient (ICC) and Bland-Altman analysis. All four algorithms systematically underestimated manually annotated global P-wave duration (bias: −8.1 to −14.4 ms; ICC=0.30-0.46). Agreement was poor to moderate across all leads for per-lead duration (ICC=0.01–0.75). P-wave positive amplitude showed good to excellent inter-algorithm agreement (ICC=0.70–0.94). P-wave terminal force V1 (PTF-V1) showed the lowest agreement, with discordance in detection and in quantitative estimates (ICC=0.64–0.93). P-wave frontal axis showed the most robust agreement (ICC=0.76–0.96, bias<3°). These findings indicate significant P-wave measurement differences and lead-dependent variability in clinically used measures.

1. Introduction

Well-established P-wave measures derived from the 12-lead electrocardiogram (ECG) are clinical markers of atrial structure and electrophysiological remodeling, used to identify atrial cardiomyopathies [1,2]. Atrial cardiomyopathy remains a recently proposed clinical entity, with diagnostic criteria still not fully defined despite its growing clinical relevance [1]. Supraventricular arrhythmias are recognised among its principal manifestations. This explains why the relationship between P-wave measures and arrhythmic risk has been extensively studied across population-based and clinical cohorts [2-4].

P-wave measures were traditionally obtained through manual caliper-based annotation of the onset and offset points on printed tracing. This method remains the reference standard in clinics, but it is time-consuming and subject to inter- and intra-clinician variability, particularly for P-waves with low amplitudes or non-standard morphologies [5]. The transition to digital electrocardiography introduced the need for a step forward to the development of automated ECG interpretation algorithms, among the most widely used being GE Marquette 12SL, Glasgow, Schiller ETM, and Philips DXL [6]. All these algorithms detect fiducial points and derive interval and amplitude measures directly from the acquired ECGs without the need for manual input [6,7]. Such algorithms underlie P-wave measures reported in clinical practice and large-scale research cohorts, making accuracy and reproducibility across manufacturers a key practical consideration. As P-wave derived measures are increasingly used across multicentre and longitudinal settings [8,9], consistency between systems determines how reliable a given measure is regardless of which algorithm produced it. However, this consistency must be established separately for each P-wave derived measure [9]. The P-wave measures of clinical interest (e.g., global and per-lead duration, positive and negative amplitude, P-wave terminal force in lead V1, and frontal P-wave axis) are each computed through their own combination of fiducial-point detection and signal-processing steps, distinct from those used for other parts of the cardiac cycle [2], [8, 9]. Several studies have compared automated measurements across ECG interpretation algorithms, although most have focused on conventional interval measurements, particularly the PR, QRS, and QT intervals [6], [7], [10,11]. Such studies, however, did not focus

specifically on the P-wave, which presents a substantially lower amplitude, variable morphology, and, in non-standard forms, far less clearly defined onset and offset boundaries than the ventricular intervals examined.

This work aims to quantify and evaluate the inter-system agreement in P-wave measures across four different algorithms: GE Marquette 12SL, Glasgow, Schiller ETM, and Philips DXL, by comparing automated outputs both against one another and against manual annotation, taken as the reference standard.

2. Materials and Methods

2.1 Dataset Selection and Manual Annotations

The dataset consisted of 140 anonymized digital 12-lead ECGs printed at standard settings (25 mm/s, 10 mm/mV) from the MUSE Cardiology Information System (GE HealthCare, Milwaukee, WI, USA). All recordings were anonymized before analysis in compliance with applicable data protection regulations. The ECG recordings were selected by two investigators (M.G. and P.G.P.), with a deliberately enriched strategy to represent the full spectrum of P-wave morphological variability encountered in clinics. Accordingly, the 140 ECGs were stratified into two major groups. The first group consisted of 40 ECGs exhibiting standard P-wave morphology in the inferior leads (II, aVF, and III). The second group consisted of 100 ECGs with non-standard P-wave morphology, including double-peaked, notched, and fractionated/multi-component P-waves in the inferior leads.

Manual annotation was performed by a biomedical engineer (M.G.) using a custom graphical user interface developed in-house in MATLAB R2025b [12], which provided interactive ECG calipers for the precise placement directly on the median beat that was extracted from the MUSE system. Annotation was first performed in the inferior leads to assess P-wave morphology, then in single-lead view with adjustable zoom for precise fiducial point placement and finally verified in the Cabrera and 12-lead layouts, with all leads sharing a common time axis, to visually identify the earliest onset and latest offset, consistent with the automated algorithms. All annotations were subsequently reviewed independently by two cardiologists (P.G.P. and F.H.), blinded to each other's assessments, with consensus reached in all cases. The final consensus annotations served as the gold standard reference for all comparisons.

2.2 Automated Algorithms and Measures

Four commercially available ECG interpretation algorithms were evaluated: GE Marquette 12SL (GE HealthCare, Milwaukee, WI, USA), Glasgow Algorithm (University of Glasgow, Scotland, United Kingdom), Schiller ETM (Schiller AG, Baar, Switzerland), and Philips DXL ECG Algorithm (Philips, Amsterdam, Netherlands). Each algorithm processed the 140 ECG recordings independently. Evaluated measures comprised global P-wave duration, frontal P-wave axis as global measures, alongside lead-specific durations, amplitude (positive and negative component) in lead I, II, III, aVF, and V1. P-wave terminal force in lead V1 (PTF-V1) was additionally extracted as a lead-specific parameter restricted to lead V1. For GE Marquette 12SL and Philips DXL, which do not report this parameter natively, it was calculated as the product of terminal P-wave duration and amplitude in V1, retaining only negative products (consistent with biphasic morphology) and reporting them as absolute values.

2.3 Statistical Analysis

Inter-algorithm agreement was assessed using the intraclass correlation coefficient (ICC, two-way mixed effects model, absolute agreement) and Bland-Altman bias analysis. ICC values were interpreted according to established thresholds: poor (<0.50), moderate (0.50-0.75), good (0.75-0.90), and excellent (≥0.90) [13]. Bland-Altman analysis was reported as mean bias and 95% limits of agreement for each algorithm pair, enabling assessment of systematic offsets and measurement dispersion [14]. For measures where manual annotation was available (global P-wave duration and PTF-V1), agreement was assessed between each algorithm and the manual reference. For measures not covered by manual annotation, agreement was evaluated between algorithm pairs only. Overall differences across algorithms for each measure were assessed using the Friedman test, followed by post-hoc pairwise comparisons with correction for multiple comparisons [15]. A p-value < 0.05 was considered statistically significant. All analyses were performed using MATLAB R2025b.

3. Results

Agreement between automated algorithms and manual annotation was assessed for global P-wave duration. All four algorithms systematically underestimated manual annotation, with mean bias ranging from −8.1 ms (Philips DXL) to −14.4 ms (GE Marquette 12SL) and 95% limits of agreement spanning approximately 50 ms across all pairs (Figure 1). ICC values ranged from 0.30 to 0.46. Among the four algorithms, Philips DXL showed the

closest agreement with manual annotation, with the highest ICC (0.46) and the smallest absolute mean bias (8.1 ms).

Inter-algorithm agreement across all P-wave measures is shown in Figure 2. For per-lead P-wave duration, agreement was poor to moderate across all leads and pairs (ICC = 0.01–0.75), with the lowest agreement observed in lead V1 and the highest in lead aVF (p-value > 0.09 for three out of four pairs). P-wave positive amplitude showed good to excellent inter-algorithm agreement across all leads (ICC = 0.70–0.94, p-value < 0.001). P-wave negative amplitude showed moderate to good agreement across leads II, III, aVF, and V1 (ICC = 0.42–0.90). P-wave negative amplitude was not reported for Lead I, as the algorithms reported near-zero values for the negative component in this lead. PTF-V1 showed the lowest overall inter-algorithm agreement, with discordance both in detection and in quantitative estimates; where detection was concordant, agreement was excellent between Schiller ETM vs Philips DXL (ICC = 0.93) and good for GE Marquette 12SL and Glasgow Algorithm (ICC = 0.86), while all other pairs showed moderate agreement (ICC = 0.64–0.65). P-wave frontal axis showed the most robust inter-algorithm agreement overall, with ICC ranging from 0.76 to 0.96 and mean bias below 3° across all pairs.

4. Discussion

This study quantified and evaluated the inter-system agreement in P-wave measures across four widely used commercial ECG interpretation algorithms, comparing automated outputs both against one another and against manual annotation as the reference standard. Inter-algorithm agreement is highly measure- and lead-dependent, with duration-based measures showing substantially lower agreement than amplitude and axis measures. Although Philips DXL showed the closest agreement with manual annotation for global P-wave duration, agreement remains limited overall, as even the highest ICC was only 0.46 and all four algorithms

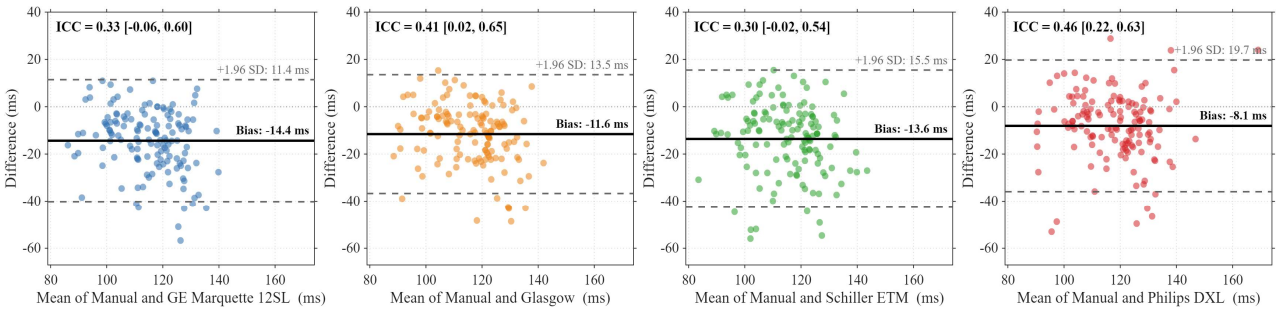

Figure 1. Bland-Altman plots comparing global P-wave duration measured by each automated algorithm with manual annotation (n = 140). In each panel, the solid line indicates the mean bias, and the dashed lines indicate the 95% limits of agreement. ICC values with 95% CI are also shown. ICC, intraclass correlation coefficient; CI, 95% confidence interval.

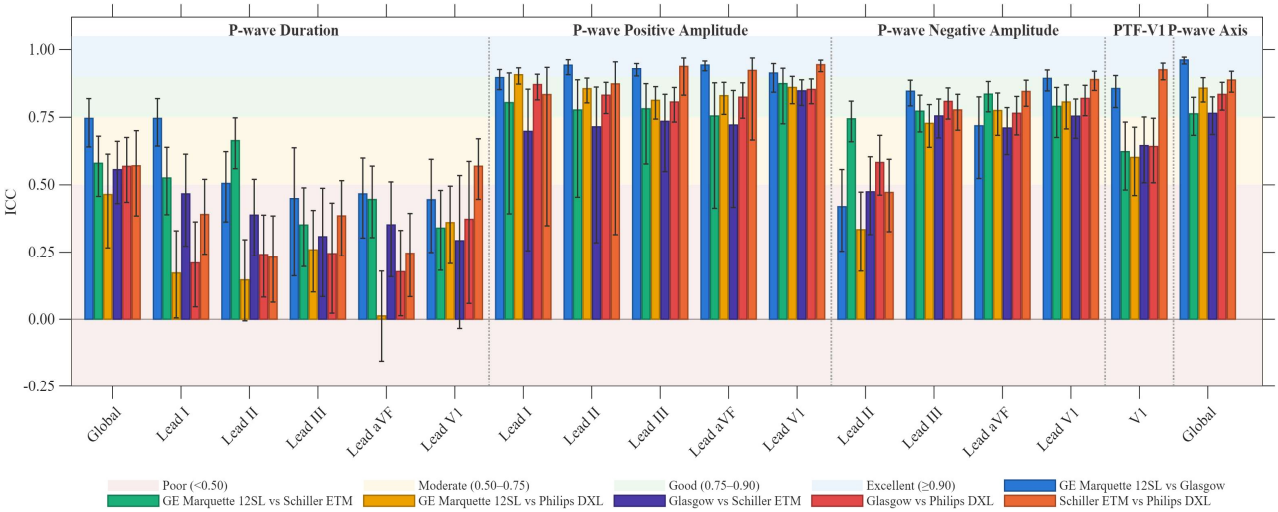

Figure 2. Pairwise inter-algorithm intraclass correlation coefficients with 95% CI for all P-wave measures across six algorithm pairs (n = 140). Error bars indicate 95% CIs. ICC, intraclass correlation coefficient; CI, 95% confidence interval.

systematically underestimated the manually annotated duration. This reflects known challenges in automated P-wave boundary detection. Amplitude also depends on onset/offset placement, but small fiducial errors affect it less. Duration, however, is directly affected by any placement error, which grows more ambiguous in low-amplitude or morphologically complex waveforms. The systematic offset observed between automated and manually annotated global P-wave duration across all algorithms is consistent with prior observations that P-wave duration shows the lowest reproducibility among standard ECG measures [5] and suggests that automated fiducial point placement differs systematically from expert annotation regardless of the platform used. This is compounded by the deliberate inclusion of both physiological (≤120 ms) and prolonged (>120 ms) P-waves, the latter corresponding to the interatrial block cut-off, which broadens the morphological range algorithms must handle. Atypical morphologies represent the cases where automated onset and offset detection is most challenged, and the poor agreement observed in leads where such morphologies are most prevalent, such as inferior leads, is consistent with this interpretation. Lead V1 showed the lowest agreement, likely reflecting its low-amplitude, biphasic morphology. PTF-V1 showed the greatest inter-algorithm discordance, consistent with its dependence on accurate detection of morphologically variable components. Conversely, measures less sensitive to waveform boundary detection, such as frontal axis, showed robust agreement across all algorithm pairs. These findings have direct implications for the use of P-wave measures in multicentre research and clinical practice, where different ECG platforms may be in use. The degree of inter-algorithm variability documented here, particularly for P-wave duration and PTF-V1, the two indices most commonly used in clinical practice, indicates that P-wave measures are not fully interchangeable across systems. A limitation of this study is that agreement was not characterized within morphological subgroups (standard vs. non-standard), which would directly quantify the contribution of morphological complexity to inter-system disagreement.

Acknowledgments

This work was supported by the Erasmus+ Programme of the European Union [2024-1-IT02-KA131-HED-000225768] and by the MUR PNRR project MCNT2 [PNRR-MCNT2-2023-12378472].

References

[1] A. Goette et al., "EHRA/HRS/APHRS/SOLAECE expert consensus on atrial cardiomyopathies: definition, characterization, and clinical implication," *Europace*, vol. 18, no. 10, pp. 1455–1490, Jul. 2016.

[2] M. A. Baturova et al., "P-wave characteristics as electrocardiographic markers of atrial abnormality in prediction of incident atrial fibrillation - The Malmö Preventive Project," *J. Electrocardiol.,* vol. 82, pp. 125–130, Dec. 2024.

[3] P. G. Platonov, "P-wave morphology: underlying mechanisms and clinical implications," *Ann Noninvasive Electrocardiol.*, vol. 17, pp. 161-169, Jul. 2012.

[4] M. A. Baturova et al., "Evolution of P-wave indices during long-term follow-up as markers of atrial substrate progression in arrhythmogenic right ventricular cardiomyopathy," *Europace*, vol. 23, pp. i29–i37, Mar. 2021.

[5] G. M. Burke, "Assessment of reproducibility--automated and digital caliper ECG measurement in the Framingham Heart Study," *J. Electrocardiol.,* vol. 47, no. 3, pp. 288–293, Jan. 2014.

[6] P. Kligfield et al. "Comparison of automated measurements of electrocardiographic intervals and durations by computer-based algorithms of digital electrocardiographs," Am Heart J.*,* vol. 167, no. 2, pp. 150–159, Feb. 2014.

[7] P. Kligfield et al. "Comparison of automated interval measurements by widely used algorithms in digital electrocardiographs," *Am Heart J.,* vol. 200, pp. 1–10, Feb. 2018.

[8] Z. Nezami et al. "Associations between biomarkers and P-wave indices in relation to atrial fibrillation development in heart failure patients," *Heart Rhythm,* vol. 23, no. 1, pp. 114–122, Jan. 2026.

[9] L. Y. Chen et al., " P Wave Parameters and Indices: A Critical Appraisal of Clinical Utility, Challenges, and Future Research-A Consensus Document Endorsed by the International Society of Electrocardiology and the International Society for Holter and Noninvasive Electrocardiology," *Circ Arrhythm Electrophysiol.,* vol. 15, no. 4, pp. e010435, Mar. 2022.

[10] J. De Bie et al., "Comparison of PR, QRS, and QT interval measurements by seven ECG interpretation programs," *J. Electrocardiol.,* vol. 63, pp. 75-82, Dec. 2020.

[11] B. Tyl et al., "Improvement and limitation of the reliability of automated QT measurement by recent algorithms," *J. Electrocardiol.,* vol. 44, no. 3, pp. 320-325, Jun. 2011.

[12] A. Sbrollini et al., "Point2ecg: An interactive software application for the identification of electrocardiographic fiducial points," *in Proc. Int. Union Phys. Eng. Sci. Med*., Singapore, pp. 12-17, Jun. 2022.

[13] T. K. Koo et al., "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research," *J. Chiropr Med.,* vol. 15, no. 2, pp. 155-163, Mar. 2016.

[14] J. M. Bland, D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *The Lancet.,* vol. 1, no. 8476, pp. 307-310, Feb. 1986.

[15] M. Friedman, "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance," *J. Am. Stat. Assoc.,* vol. 32, no. 200, pp. 675-701, Dec. 1937.

Address for correspondence:

Pyotr G Platonov
Department of Cardiology, Lund University, Lund, Sweden,
pyotr.platonov@gmail.com