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

Meri Gjika*, Agnese Sbrollini, Anders Björkelund, Claus Graff, Dillon J Dzikowicz, Fredrik Holmqvist, Junmo An, Peter W Macfarlane, Ramun Shmid, Laura Burattini, Pyotr G Platonov

Università Politecnica delle Marche, Ancona, Italy

Introduction: Well-established P-wave measures derived from the 12-lead electrocardiogram (ECG) represent clinical markers of atrial structure and electrophysiological remodelling, used in the identification of atrial cardiomyopathies. However, automated ECG interpretation algorithms differ in the quantitative characterisation of the P-wave. This work aimed 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.

Methods: Data consisted of 140 anonymised 12-lead ECGs with enriched P-wave morphological variability, independently processed by all four algorithms. It comprised 40 ECGs with standard P-wave morphology and 100 with non-standard morphology (i.e., double peaked, notched, fractionated/multi-component). Manual annotations served as the gold standard, obtained by a biomedical engineer using custom ECG calipers and refined by two cardiologists, who independently reviewed and adjusted the onset and offset threshold positions where needed. Agreement was assessed using the intraclass correlation coefficient (ICC) and Bland-Altman bias analysis.

Results: For global P-wave duration, GE Marquette 12SL and Glasgow showed near-identical results (ICC=0.998, bias=-0.06 ms), though all algorithms underestimated manual measurements by 8-14 ms (ICC =0.30-0.46). In leads I and II, all algorithms underestimated by 11-17 ms, with ICC (0.21-0.41) and moderate inter-algorithm agreement (ICC =0.17-0.75). Per-lead duration agreement was lowest in V1 and highest in aVF, where three out of four algorithms converged ($p>0.09$). P-wave amplitude showed high inter-algorithm agreement for the positive component (ICC=0.70-0.94), consistently across leads I and II (ICC=0.70-0.91 and 0.71-0.94, respectively). P-wave terminal force in V1showed the lowest overall agreement, with discordance both in detection and in quantitative estimates. P-wave axis had the most robust inter-algorithm agreement overall (ICC=0.76-0.96 across all pairs, bias <3°).

Conclusion: These findings indicate significant P-wave measurements differences and lead-dependent variability in measures widely used in clinics.