

Repeat variation and definition dependence of ECG/VCG markers: an exploratory audit of one implementation (LoopLens 10.1)

Leo Probst

`leo.probst.research@proton.me`

Johannes Kepler University of Linz <https://orcid.org/0009-0003-4130-7981>

Research Article

Keywords: electrocardiography, vectorcardiography, MIMIC-IV-ECG, measurement validity, reproducibility, reconstruction, delineation

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11410667/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

Objective

Physiological names do not fully determine the measurement definition of an electrocardiogram (ECG) or vectorcardiogram (VCG) marker. We examined how the values of one implementation (LoopLens 10.1) change between routine recordings 0–2 h apart, between two reconstruction matrices and between boundary references, in a catalog of 2,005 markers.

Approach:

Development-data audits in MIMIC-IV-ECG (15,693 pairs; 5,648 recordings for the matrix comparison), expert-boundary comparison in LUDB, repeat pairs from PTB-XL and public randomized QT controls. Replication was prespecified in a person-disjoint test partition, previously opened for other analyses.

Main results:

The median ICC(A,1) of 1,958 estimable non-circular markers was 0.468 (311 reached 0.75); an ICC cannot separate error, change and low between-person spread. The repeatability coefficient (RC, 1.96 SD of paired differences, including true change) was 27.27 ms for QRS duration, 61.78 ms for QT corrected by the Fridericia formula (QTcF) and 82.41° for the QRS-T area angle. Under author-fixed rules, replacing the Kors with the inverse Dower matrix gave 1,045 dependent, 684 conditional and no stable outputs among 1,729 non-identical estimable non-circular ones, yet ± 4 -ms boundary shifts alone gave 496 dependent of 1,724; 29 circular markers were not classified. Against LUDB experts (amplitudes normalized per lead), the median P-onset error was + 14 ms with extreme-boundary and + 1 ms with lead-median references. PTB-XL repeat variation (pairs up to 24 h apart) had equal or lower point estimates than MIMIC (R 0.77–0.99), with no class decided. In public randomized recordings, the production (standard LoopLens) and source-annotation routes differed in 7 of 16 estimable treatment contrasts (exploratory, unadjusted). In the test partition, the historical matrix classification and P availability replicated; repeatability criteria were not met, with higher test ICCs.

Significance

These are properties of one implementation. They do not show general failure of VCG markers, certify clinical interchangeability or set treatment thresholds.

Key findings

- Repeat variation between routine ECGs is large and cannot be split into error and change. In 15,693 pairs recorded 0-2 h apart, the median ICC over 1,958 non-circular markers was 0.47 and 16%

reached 0.75. RC was 27 ms for QRS duration, 62 ms for QTcF and 82° for the QRS-T area angle. These describe repeat variation including true change; they are not measurement-error estimates. Machine-reported measurements of the same recordings varied as much for QRS duration (device-to-LoopLens RC ratio 0.97) and less for QTcF (0.81; post hoc, S3h). Substantial repeat variation was also observed in the machine-reported measurements; these comparisons do not identify a shared variance component or its causes.

- The reconstruction matrix matters for many outputs. Among 1,729 non-identical estimable non-circular outputs, switching from Kors to inverse Dower gave 1,045 classed dependent, 684 conditional and none stable (author-fixed rules; identical outputs excluded). For scale, the same rules labelled 496 of 1,724 outputs dependent (295 stable) after ± 4 -ms boundary shifts with an unchanged matrix, and 1,554 of 1,721 estimable catalog outputs dependent (none stable) at identical recorded time (pooled, descriptive; S3f). The median half-width of the matrix limits of agreement was 0.54 of the repeat variation RC (S2g), so the label reflects the strictness of the rules as much as the matrix. Feeding the other matrix to the delineator changed QRS duration and QT only slightly (limits of agreement about ± 1 and ± 5 ms) and P-wave outputs more (availability and about ± 22 to 27 ms).
- Boundary errors depend on the reference. The median P-onset error against LUDB experts was +14 ms (mean absolute error 31.9 ms) against extreme-boundary references and +1 ms (23.7 ms) against lead medians.
- Elsewhere the picture is partial. In PTB-XL pairs up to 24 h apart, point estimates of repeat variation were equal or lower than in MIMIC pairs 0-2 h apart (QTcF R 0.99; QRS duration 0.77; area angle 0.85 [0.74, 0.95]); no primary class was decided. On identical PTB-XL pairs, BRAVEHEART had the lower repeat variation for the area angle and two area-based markers (repeatability, not accuracy; S14d).
- For clinical readers: no clinical threshold follows. The next step is independent software on the same recordings, blinded boundary marking and repeat pairs with documented stable clinical state.

Evidence status

Exploratory measurement audit on development data, with prespecified internal replication in a person-disjoint test partition of the same database, previously opened for other exploratory analyses. The historical matrix classification (the catalog has 29 circular markers, 26 of them in the historical set of 1,755 outputs) and P availability of the release output replicated; the repeatability criteria were not met (S13). Analyses added after the first draft (route differences, proxy sensitivity, circular markers, replication composition, the comparison with machine-reported measurements) are post hoc and marked as such. The PTB-XL repeat-pair analysis was also added after the first draft but specified before the PTB-XL values were read; its identical-pair comparison with BRAVEHEART and the QTcF band-edge recomputation are post hoc. LUDB and PTB-XL were used earlier in LoopLens development, so neither provides independent validation; LUDB amplitudes are normalized per lead.

1. Introduction

Physiological names do not fully determine the measurement definition of an ECG/VCG marker. A value depends on its formula, its baseline and origin choices, its boundary decisions and, for spatial quantities, the reconstruction matrix. We examine how the values of one implementation change between routine recordings, reconstruction matrices and boundary references. A physiological name or disease association does not establish measurement validity.

Earlier work addressed same-day ECG reproducibility (Haq *et al* 2021), reproducibility across segments of one recording (Perez-Alday *et al* 2018), angle formulations and heart rate (Andršová *et al* 2022), Frank reconstruction (Edenbrandt and Pahlm 1988, Kors *et al* 1990, Schreurs *et al* 2010), delineation sensitivity (Young *et al* 2021) and automated P-wave analysis (Platonov 2026). De la Garza Salazar and Egenriether (2024) analyzed 315 Kors-derived parameters in 664 adults, and Jekova *et al* (2016) studied biometric verification more than 1 year apart. BRAVEHEART is an open-source package for automated ECG/VCG analysis (Stabenau and Waks 2023). None of these sources validates our catalog, and we do not claim priority.

We ask three questions about one implementation: how much do routine ECG/VCG markers vary between recordings 0–2 h apart, how far do they depend on the reconstruction matrix and delineation, and how do boundary errors depend on the reference definition? Disease-association and model-validation questions are outside the scope.

2. Methods

2.1. Data, design and replication

MIMIC-IV-ECG and linked MIMIC-IV routine clinical information supplied the retrospective development data: routine diagnostic ECGs acquired at Beth Israel Deaconess Medical Center, Boston, in 2008–2019 with devices from various manufacturers, 12 leads, 10 s, 500 Hz (Gow *et al* 2023). The acquisition period and device of the analyzed ECGs were not established from the local metadata (S10a). Training and validation are two internal development partitions, already inspected (cross-fitting exchanges their roles). PTB Diagnostic (Bousseljot *et al* 1995, 2004), PTB-XL (Wagner *et al* 2020, 2022) and the Lobachevsky University Electrocardiography Database (LUDB; Kalyakulina *et al* 2020, 2021) were used for external technical comparisons; none provides independent disease-model validation, and LUDB and PTB-XL were used earlier in LoopLens development.

The measurement-replication protocol was fixed on 12 October 2026, before newly extracted test marker values were inspected. Blinding here is question-specific: when the specification was fixed, neither the author nor the coordinating AI-assisted workflow had seen test results for the repeatability, matrix or P-availability questions, and an AI-assisted analysis workflow then computed the test statistics once, following the fixed specification; no human analyst other than the author took part, so this is not independent analyst blinding. The person-disjoint test partition had already been opened for other

research questions. Two markers that failed development-only extraction equivalence and three previously computed matrix markers were excluded in advance (S13); sensitivities cannot rescue a failed criterion.

Denominators (ECGs, persons, pairs, markers) are separate and overlapping branches are not added; the repeat-branch flow is in S5a.

2.2. Implementation

Measurements came from LoopLens (version 10.1; release of 7 October 2026), an in-house Python pipeline for ECG delineation, VCG reconstruction and marker calculation; the author is its developer. The MIMIC-IV-ECG runs used a runtime snapshot that differs from the release folder in seven files of later analysis stages; the code for signal processing, delineation, reconstruction and marker calculation is identical (S10a). Diagnostic signals were filtered with a zero-phase third-order Butterworth filter (0.05–150 Hz) and delineation signals with a fourth-order filter (0.5–40 Hz) plus 2-s running-median baseline correction. Multilead R detection precedes fusion of NeuroKit2 (Makowski et al 2021, version 0.2.13), multilead and Kors-XYZ delineation candidates (nominal weights 0.40/0.35/0.25); a representative fused beat supplies boundaries. A delineation-quality score (clipped to [0,1]) is an implementation index, not validated signal quality. The primary spatial angle uses origin-corrected QRS and J-to-T-end area vectors. Formula, unit, segment, origin and software version jointly define a marker (S1, Supplementary Data S2). Kors denotes the Kors regression matrix (Kors et al 1990); the production route is this standard processing chain, as opposed to the source-annotation route of section 2.7. The catalog of 2,005 identifiers is an inventory of marker outputs; related identifiers differ in formula, segment or normalization (S1, S3d). A training-data screen ($|\text{Spearman}| > 0.999$) left 1,883 components, 111 of them with more than one member (median ICC of one representative per component 0.459). P availability is the fraction of ECGs for which the release returned a finite P duration from its stored output (the release cache). A recomputation can return a value where the stored output has none: 932/1,500 (62.1%) versus 1,354/1,500 (90.3%) in a reference sample of persons with normal sinus ECG reports, so a missing stored value does not mean that no P wave is measurable (S3a, S5).

2.3. Repeat variation

QTcF is the QT interval corrected by the Fridericia formula ($QT/RR^{1/3}$, RR in s). Clinical pairs were adjacent ECGs from the same admission and setting, 0–2 h apart, with successful extraction and quality ≥ 0.6 . The report-defined stability proxy also required a positive gap, machine-reported sinus rhythm, an equal QRS-duration class and no pacing report; it does not prove physiological stability. ICC(A,1) is two-way, single-measurement and absolute-agreement (McGraw and Wong 1996); a marker was estimable with ≥ 30 persons and a positive ANOVA denominator. Intervals are 95% percentile intervals from 1,000 person-level bootstrap draws (seed 77010).

RC is 1.96 times the SD of the centered paired differences, in the marker's unit. It is a descriptive spread of repeat differences. It contains true change between recordings and selection effects and may understate error if errors are correlated between recordings; it is neither a standard error of

measurement nor a smallest detectable change, so we call it repeat variation. Bland-Altman limits (Bland and Altman 1986) share it. The $ICC \geq 0.75$ line applies the “good reliability” boundary of Koo and Li (2016) to point estimates; those authors interpret ICC by its 95% confidence interval. As a further post hoc comparison, the machine-reported measurements stored with MIMIC-IV-ECG (QRS duration, QT, QTc derived by the Fridericia and Bazett formulas, P duration, PR interval, frontal QRS-T angle) were analyzed on the same pairs with the same estimators; the rule was written to a file before the run (S3h).

Markers that are angles on a circle need circular differences; a linear ICC, an SD-based RC and a matrix class computed from raw angle differences are not valid for them. Periods were taken from the code definition of each marker (S2i). These 29 markers are listed as not classified (periodic) in the main results and are excluded from the catalog summaries and matrix counts; the historical values including them are reported as a sensitivity (S2i). Their repeat variation is described by the 95th percentile of the shortest angular distance (D95), which is a quantile and not comparable to the SD-based RC.

2.4. Matrix and delineator-input dependence

The enriched catalog sample used one ECG per person with identical delineation under both matrices, so the comparison isolates the reconstruction used for marker calculation. Differences are inverse Dower minus Kors. Each fold required ≥ 100 finite pairs. Technical stability required Spearman ≥ 0.95 , relative limits-of-agreement width ≤ 0.5 and quintile switches ≤ 0.10 ; dependence required Spearman < 0.80 , width > 2.0 or switches > 0.40 ; otherwise conditional; the worse fold decides (width: the width of the limits of agreement relative to the SD of the values, S2f; quintile switches: the share of ECGs moving to another fifth of the ranking). These are author-fixed technical rules, not clinical tolerances or an equivalence test. Because fiducials were shared, native outputs are identical under both matrices at those fiducials. A prespecified sensitivity analysis (more than 490 development ECGs) supplied inverse-Dower instead of Kors XYZ to the delineator itself and is reported marker by marker (S2h). The 29 circular markers are not classified (section 2.3). Further technical sensitivities (± 4 -ms boundary shifts, affine conversion, resampling) used the historical set (S2g). The $> 105^\circ$ area-angle label is a literature-motivated stress test, not a clinical cutoff.

2.5. Boundaries and Frank comparison

LUDB 1.0.1 has 200 ten-second, twelve-lead recordings at 500 Hz; the record headers give each lead its own gain factor so that every lead spans exactly 1 mV, which the dataset documentation does not describe, so boundary errors (production minus expert) are specific to this operating point. Errors were computed against the extreme-boundary expert reference (earliest onset, latest offset) and, post hoc, against lead medians; neither matches production fusion exactly. PTB provided measured Frank leads for the comparison of reconstructed and measured spatial angles; its fiducials are partly Kors-assisted, so no independent gold standard exists (S2). A single-author visual screen of P-wave boundaries is described in S3a.

2.6. External repeat pairs

The PTB-XL analysis, added after the first draft but specified before the PTB-XL values were read, used pairs of recordings of the same person up to 24 h apart (312 pairs, 259 persons). R is the PTB-XL RC divided by the MIMIC RC, with a combined interval from person-cluster bootstrap (2,000 draws) and a lognormal uncertainty of the MIMIC RC; the band 0.80–1.25 defined the classes concordant (interval inside), lower (below) and undecided (overlapping an edge). A later 20,000-draw recomputation addressed Monte Carlo error at the QTcF band edge. LoopLens is the author's own tool and PTB-XL was used in its development and calibration; repeatability is not accuracy (S14). The post hoc comparison with BRAVEHEART on identical pair sets used BRAVEHEART 1.8.2 outputs from the separate benchmark run (compiled batch application, unchanged default settings, median beat; QTcF derived with the Fridericia formula; S14d).

2.7. Public QT controls

ECGRDVQ randomized crossover recordings (Johannesen et al 2014, Vicente et al 2015; ECGRDVQ 2016) were processed with the production route (the standard LoopLens 10.1 chain) after resampling to 500 Hz. The fixed primary endpoint was placebo- and baseline-corrected QTcF change at 2 h (Holm correction over four tests), with prolongation expected for dofetilide and quinidine, a smaller increase for ranolazine and none for verapamil. Prior awareness: the first seven rows of an earlier parallel 2-h table were displayed during the author's own extraction, earlier than the protocol foresaw, and the source-route adapter was defined after early single production outputs had been seen; no marker, time point or test was changed, and this is an independent technical recomputation with disclosed knowledge of results, not an analyst-blinded confirmation (S6b). Source measurements (annotated QT with recording RR, fixed pre-P baseline) form a second route. As a post hoc analysis, we estimated for each drug and marker the participant-level difference D between the production and source double differences, with 95% participant-bootstrap intervals (10,000 draws). The secondary family comprised four drugs by four markers (area angle, Tpeak-Tend, QRS duration, normalized Fréchet distance); with the four QTcF contrasts this gave 20 contrasts, of which 16 were estimable because the source annotations lack T onset for the four Fréchet contrasts (S6b).

2.8. Reporting and AI use

GRRAS, STROBE and RECORD guided reporting (Kottner et al 2011, von Elm et al 2007, Benchimol et al 2015); remaining gaps are listed in S8c. Intervals are pointwise. Counts of 1–9 persons are shown as < 10. Age and administratively recorded sex are reported by population (S5b); gender identity was undocumented, and sex- and age-stratified repeat variation is descriptive (S3e). The author used generative AI tools (Anthropic Claude and OpenAI Codex) to assist with analysis code, literature searches, data analysis workflows, drafting and editing of text, and figure preparation. AI-assisted output was reviewed, corrected where necessary and approved by the author, who takes full responsibility for the content. Patient-level analyses were run locally. AI tools are not listed as authors.

3. Results

3.1. Where each result comes from

Table 1 lists the source, generalization status and calculation state of each core finding.

Table 1

Source, generalization status and calculation state of the three core findings. All analyses used LoopLens 10.1 (release of 7 October 2026; section 2.2).

Core finding	MIMIC-IV-ECG (development)	Other data	Status	Calculation state
1. Repeat variation of clinical pairs	15,693 pairs, 10,187 persons; test partition 3,897 pairs	PTB-XL 312 pairs, 259 persons	Internal replication criteria not met; external comparison: equal or lower point estimates, no class decided	A separately written aggregation script reproduced the estimates from the pair-level values; historical generator identity incompletely documented
2. Matrix and delineator dependence	5,648 ECGs; historical classification replicated in test partition	PTB (Frank leads, 290 persons); LUDB	Historical classification (with circular markers) replicated internally; non-circular main classification not replicated separately; no external catalog test	Class counts reproduced; same documentation limit
3. Boundary errors depend on reference	Not applicable	LUDB only (200 recordings, normalized amplitudes, used in development)	Not transferable to physical amplitudes; no independent validation	Rerun reproduced all 14,631 landmark rows; historical runtime copy inaccessible
Added after the first draft	Proxy, circular markers, replication composition and machine-reported measurements (post hoc)	ECGRDVQ route differences (post hoc); PTB-XL pairs (specified before the values were read)	Exploratory	Specifications fixed and reproduction criteria passed before the run

3.2. Short-interval repeat variation

Among 15,693 pairs from 10,187 persons, the median ICC over the 1,958 estimable non-circular markers was 0.468 (311 reached 0.75; 15.88%); in the stability proxy (4,596 pairs, 3,560 persons) it was 0.526 over 1,946 markers (482; 24.77%). Intervals for these reduced sets were not recomputed; the contrasts

and intervals below refer to the historical set of all 1,987 estimable markers, which includes the 29 circular markers (median 0.4665 [0.4591, 0.4778]; S0a). Table 2 and Fig. 1 give the key measures (none is circular); Fig. 2 shows the distribution of the marker-level ICC; Supplementary Data S2 holds all markers.

Table 2

ICC(A,1), bias and repeat variation (RC) of key measures in all clinical pairs, with the stability proxy for comparison. Intervals are 95% person-bootstrap percentile intervals.

Measure	Finite pairs / persons	ICC [95% interval]	Bias	RC [95% interval]	Proxy ICC	Proxy RC
P duration, multilead fusion (ms)	3,397 / 2,731	0.4915 [0.4431, 0.5382]	0.29	47.84 [44.71, 50.97]	0.4539	50.94
QRS duration (ms)	15,125 / 9,962	0.8665 [0.8523, 0.8794]	0.22	27.27 [25.86, 28.77]	0.9169	15.99
QTcF (ms)	12,454 / 8,629	0.7151 [0.6969, 0.7343]	1.98	61.78 [59.48, 63.99]	0.7894	47.55
Spatial area QRS-T angle (°)	13,312 / 9,115	0.6170 [0.6031, 0.6321]	0.13	82.41 [80.85, 83.91]	0.6934	71.06

Table note: Proxy intervals are in S3. Without the equal-QRS-class condition the proxy QRS RC was 17.36 ms [16.11, 18.61]; the other three RCs moved by less than 1 unit.

At identical recorded times, the area-angle RC was 81.46° [71.86, 89.72] (294 pairs), so elapsed time alone does not explain it; in report-normal pairs it was 46.64° [40.30, 52.74] (322 pairs).

Machine-reported measurements (post hoc). On pairs with values from both tools, the machine-reported QRS duration varied as much as the LoopLens value (RC 26.17 versus 26.98 ms; RC ratio, device over LoopLens, 0.97 [0.91, 1.03]) and the QT interval slightly more (1.06 [1.03, 1.09]). Derived QTcF (0.81 [0.78, 0.85]), P duration, PR interval and the frontal QRS-T angle (0.58, 0.48 and 0.67) varied less, although the device P onset equals 40 ms in half of the recordings and the angle is not the spatial area angle (S3h). The device measurements showed similar repeat variation for QRS duration and QT, so this variation is not specific to LoopLens (a descriptive observation, not an identified shared cause); for QTcF and the frontal angle the device values varied less, so a tool-specific component cannot be excluded.

Removing the QRS-class condition left the catalog median unchanged (0.5237 versus 0.5242), but the proxy-minus-all contrast of 0.058 fell to 0.024 [0.013, 0.035] with persons held fixed and to 0.010 [0.002, 0.021] with equal person weights as well (post hoc; S3g), so proxy values are not an isolated stability result.

For the 26 directed circular markers, D95 was 92–166° (median 153°), close to 143–172° when the same values were paired at random, although the median angular distance was 6–61° (S2i). Including the 29 circular markers changes the catalog median ICC by 0.001 and the dependent share of matrix outputs by 0.6 percentage points (post hoc).

PTB-XL repeat pairs. In PTB-XL pairs up to 24 h apart the point estimates of repeat variation were equal (QTcF, R 0.99) or lower (QRS duration 0.77 [0.58, 1.01]; area angle 0.85 [0.74, 0.95]) than in MIMIC-IV-ECG pairs 0–2 h apart (Table 3); no primary class was decided (all were undecided). For QTcF the lower limit sat on the band edge (0.798 in the original run, 0.802 pooled over 20,000 draws), so the recomputation yielded no robust class. The QRS RC was driven by a few pairs, including 12 with pacemaker-flagged recordings (R 0.65 [0.54, 0.77] without them; S14b). On identical marker-wise pair sets (post hoc), BRAVEHEART had the lower repeat variation for the area angle (RC 50.09 versus 69.65°; RC ratio BRAVEHEART over LoopLens 0.72 [0.57, 0.88]), the QRS area magnitude (0.78 [0.50, 0.89]) and the sum of absolute QRS-T integrals (0.83 [0.61, 0.96]), the three markers whose intervals excluded 1; its point estimate was also lower for four of the other six markers, whereas LoopLens had the lower point estimate for QTcF (1.32 [0.95, 1.82]) and QT (1.14 [0.99, 1.35]). This compares repeatability, not accuracy (S14d).

Table 3
Repeat variation in 24-h PTB-XL pairs relative to MIMIC-IV-ECG ($R = RC \text{ PTB-XL} / RC \text{ MIMIC}$; band 0.80–1.25).

Marker	Persons	RC PTB-XL [95% interval]	RC MIMIC	R [combined 95% interval]	Class
QRS duration (ms)	259	21.06 [15.93, 27.03]	27.27	0.77 [0.58, 1.01]	undecided
QTcF (ms)	223	60.86 [49.64, 72.10]	61.78	0.99 [0.80, 1.17]	undecided (band edge)
Area QRS-T angle (°)	232	69.84 [61.02, 78.16]	82.41	0.85 [0.74, 0.95]	descriptive (availability rule differs)

Table note: QT interval and peak angle (undecided) are in S14; 13 of 2,005 identifiers were covered. Pair interval, devices (Schiller) and population differ from MIMIC and cannot be separated; LoopLens was used in PTB-XL development.

3.3. Matrix and delineator dependence

On the fixed 5,648-record sample, 5,307 finite area-angle pairs showed Spearman 0.909, bias + 9.593° and limits of agreement – 30.312° to + 49.498°. Substituting inverse Dower for Kors switched the > 105° label in 710 of 5,307 pairs (13.4%). Table 4 gives the catalog classification (the bootstrap intervals in S2f are historical intervals of the historical set, not intervals of the adjusted main count; the historical composition is in Figure S1).

Table 4

Full-catalog same-ECG comparison of inverse Dower and Kors reconstruction (technical classification).

Quantity	Count	Meaning
Enriched development ECGs	5,648 (one per person; 2,824 per fold)	Not prevalence sampling
Circular catalog markers	29, not classified (periodic)	Raw-angle class not valid; values in S2i
Non-identical estimable non-circular outputs	1,729	Excludes identical and constant entries
Dependent	1,045 (60.4%)	Author-fixed technical criteria
Conditional	684 (39.6%)	Intermediate
Stable	0	Not an equivalence test
Identical nonconstant outputs	214 (+ 1 constant)	Shared, partly Kors-assisted fiducials
Other catalog markers	31 not estimable in the full 5,648-ECG comparison and 1 not testable	Not the 31 of the boundary benchmark below

Table note: Percentages are of the 1,729 non-identical estimable outputs. The historical set (1,755 outputs; 1,071 dependent, 61.0%, and 684 conditional) contains 26 of the 29 circular markers, all dependent; the other three were identical under both matrices and lie among the 217 identical nonconstant outputs of the historical count (214 without them).

In the ± 4 -ms Kors/Kors boundary-shift experiment (500 ECGs; historical set of 1,755 outputs, which contains 26 of the 29 circular markers), 295 outputs were stable, 933 conditional and 496 dependent among the 1,724 estimable in that sample. Applied to repeat recordings at identical recorded time (360 pairs), the same rules labelled 1,554 of 1,721 pooled estimable catalog outputs (circular markers included) dependent and none stable (descriptive; S3f), and the median ratio of the matrix limits-of-agreement half-width to the repeat variation RC was 0.54 (IQR 0.34–0.78; S2g). The label “dependent” thus means exceeding author-fixed rules, which many outputs do under repeat recordings or small boundary shifts as well; the dependent share ranged from 40.7% of amplitudes and areas to 92.5% of counters and flags (S3d). The 31 outputs not estimable in the boundary experiment are a different group from the 31 catalog markers not estimable in the full comparison (Table [4](#); S2g-4). This experiment shows that the classes are sensitive to small boundary shifts and to the classification rules; it does not decompose matrix dependence at fixed boundaries. Affine calibration is in S2g.

Feeding inverse-Dower instead of Kors XYZ to the delineator (more than 490 ECGs) changed QRS duration and QT interval within -1.3 to 1.2 ms and -4.8 to 4.8 ms (limits of agreement), similarly in LUDB and PTB, whereas PR interval and P duration were missing from one of the two delineations in about 5.5% of ECGs (more than 490 evaluated) and changed by up to about ± 27 ms where both were

available. We therefore report by marker: QRS duration and QT changed little under the substitutions examined, PR and P duration changed more (S2h).

3.4. Boundary errors against experts

In LUDB, 1,831 of 1,869 expert QRS groups were matched; P-onset errors were available for 1,250 of 1,402 attached P references and T-offset errors for 1,642 of 1,644 (Table 5). With lead-median references the median P-onset error changed from + 14 to + 1 ms and the T-offset error from – 14 to + 4 ms, but the QRS-offset median moved from – 2 to + 12 ms and the T-onset median from + 6 to – 26 ms, so the lead-median reference was not uniformly favorable. At P onset the mean absolute error fell only from 31.9 to 23.7 ms, so much of the error remained under either reference.

Table 5

LUDB boundary errors (production minus expert) against the extreme-boundary reference and, post hoc, against lead medians. Counts are conditional on matched QRS groups; intervals are 95% recording-cluster bootstrap intervals. Amplitudes are normalized per lead.

Boundary	Pairs / reference rows	Mean ± SD (ms)	Median [95% CI]	MAE [95% CI]	Median, lead-median reference	MAE, lead-median reference
P onset	1,250 / 1,402	6.32 ± 48.01	14 [12, 16]	31.88 [29.14, 34.76]	1	23.70
P offset	1,250 / 1,402	–12.24 ± 22.82	–10 [– 10, – 8]	14.75 [13.21, 16.34]	4	10.98
QRS onset	1,831 / 1,831	4.25 ± 13.52	6 [4, 6]	10.27 [9.37, 11.28]	–5	9.01
QRS offset	1,831 / 1,831	–2.84 ± 11.59	–2 [– 2, – 2]	8.07 [7.36, 8.81]	12	13.33
T onset	1,642 / 1,644	6.06 ± 29.74	6 [4, 10]	21.52 [19.66, 23.73]	–26	33.95
T offset	1,642 / 1,644	–14.65 ± 32.80	–14 [– 16, – 12]	23.55 [20.93, 26.23]	4	16.92

Complete 150-ms matching against the extreme references gave P-onset sensitivity 86.74% [83.54, 89.52] and T-offset sensitivity 91.87% [89.46, 94.35] (S3b; predictive values in S0a). These values describe annotation definitions, not clinical performance.

A single-author visual screen of long P waves in a report-selected sample (S3a) yields counts of the author’s own judgments, made with the proposed boundary lines visible, by one rater and after a change of reference and tolerance; they are not error prevalence. P duration was missing from the release cache

in 568 of 1,500 reference persons (37.9%); a recomputation recovered 422 of them, with longer durations than in the cache-available group (P97.5 240 versus 181 ms), and 146 (9.7%) remained without a value (S3a).

3.5. Frank comparison

Integral-angle deviations (reconstruction minus Frank) depended on the population. In 51 PTB-healthy persons, inverse-Dower bias was + 15.51° [10.93, 20.12] and Kors bias – 1.00° [– 4.537, 2.421]; the P95 of absolute differences was 40.64° [30.28, 53.38] and 24.03° [17.54, 35.20]. Agreement of the net rotation direction (clockwise or counterclockwise) between the Kors-reconstructed and the measured Frank T-wave loop was 76.47% in the transverse plane, below the prespecified 90% agreement criterion, whereas the left-sagittal plane passed arithmetically; neither implies transport to individual patients (S2d). The frontal minor-area loop class depended strongly on its threshold (S2a-2, S0a).

3.6. Public randomized QT controls

At 2 h, production delineation recovered the direction of all three expected positive QTcF effects after Holm correction (Table 6). The two routes use the same recordings and are not independent replications.

Table 6
QTcF at 2 h (ECGRDVQ), production and source-annotation routes, and their participant-level difference D (post hoc). ΔΔ: placebo- and baseline-corrected change.

Drug	Participants	Production ΔΔQTcF, ms [95% CI]	Source ΔΔQTcF, ms [95% CI]	D, production minus source [95% CI]
Dofetilide	22	61.96 [55.05, 69.33]	62.28 [53.64, 72.44]	–0.32 [– 5.30, 3.75]
Quinidine	21	66.86 [60.60, 73.52]	78.37 [70.41, 86.34]	–11.51 [– 16.63, – 6.47]
Ranolazine	22	5.78 [2.35, 9.20]	7.39 [3.73, 11.20]	–1.61 [– 3.99, 0.60]
Verapamil	22	–0.48 [– 5.12, 4.26]	3.42 [– 0.50, 7.15]	–3.90 [– 7.95, – 0.33]

For orientation, Vicente et al (2015, Table 2) report concentration-model QTcF changes at the mean maximum concentration of + 73.6 ms (dofetilide), + 78.9 ms (quinidine) and + 12.0 ms (ranolazine); these are not time-matched to our 2-h estimates. Ours were lower for dofetilide (61.96 and 62.28 ms in the production and source routes) and ranolazine (5.78 and 7.39 ms); for quinidine the source-route estimate (78.37 ms) was numerically close to the published value, and the production route (66.86 ms) was 11.5 ms lower than our source-annotation estimate. Recovery of direction therefore does not imply recovery of magnitude.

The two routes did not estimate the same treatment effects for all markers. Of 16 estimable contrasts (4 QTcF, 12 secondary), 7 had pointwise 95% bootstrap intervals for D that excluded zero, an exploratory finding without multiplicity adjustment. D was largest for Tpeak-Tend under quinidine (– 15.65 ms [– 20.65, – 10.92]); production QRS duration rose under dofetilide (D = + 3.56 ms [0.89, 6.26]) and quinidine

(+ 3.81 ms [1.29, 6.52]), whereas the annotated route did not (participant-level correlation of QRS values 0.03–0.23). The routes can thus disagree in a treatment-dependent way; the estimates do not show which is closer to the true effect, because boundary definition, origin and RR pairing differ. Production area-angle and normalized Fréchet responses to dofetilide and quinidine were unsupported after multiplicity correction (S6b-3).

In a routine-care control panel of six expected positive and four negative control relations in the validation partition, three positive controls were recovered (expected direction, $p \leq 0.01$) and one negative control was flagged (AUC differing from 0.5 by at least 0.05); computation flags are not clinical verification (S6-1).

3.7. Test-partition replication

The historical matrix classification and P availability replicated under their prespecified technical criteria. The matrix criteria were fixed on the historical classification, which includes the circular markers, and were not recomputed without them. Class agreement was 93.7% [91.8, 94.5], κ 0.87 [0.82, 0.88], the dependent fraction $1,074/2,000 = 0.537$ versus $1,068/2,000 = 0.534$ in development, and P availability of the release output $921/1,500$ versus $932/1,500$.

Repeatability criteria (also on the historical marker set including circular markers) were not met in either population. Across 3,897 pairs from 2,549 persons, the test median ICC was 0.507 [0.489, 0.537] over 1,979 markers against a rebased development benchmark of 0.467; in the proxy (1,174 pairs, 931 persons) it was 0.559 [0.545, 0.583] against 0.525, with an RC-band share of 79.7%, below the required 80%. The criteria failed because the upper interval limits exceeded the two-sided bands, not because reproducibility was lower: the test-minus-development difference was + 0.040 [0.014, 0.066], and 74% of markers had a higher test ICC. Standardizing the development pairs to the test case mix left the median nearly unchanged (S0a). A post hoc, model-dependent decomposition did not point to a higher within-person variation in the test partition; age, sex, device and the measured spread of the test pairs were not analyzed (S13e, S0a). The criterion decisions are unchanged; this was internal replication, not external validation.

4. Discussion

4.1. Three questions to keep apart

For a given measurement implementation, clinical repeat variation, reconstruction dependence and reference dependence need to be judged separately. The catalog median ICC for ECGs recorded 0–2 h apart was moderate (ICC 0.87 for QRS duration, 0.49 for P duration), but this repeat variation includes clinical change, and the shares of definition, processing and state were not separated. Matrix substitution moved 1,045 of 1,729 comparable non-circular outputs (60.4%) into our technical dependence class and none met the stability rules; the same rules also labelled 496 of 1,724 outputs dependent after ± 4 -ms boundary shifts without any change of matrix, so the label is a statement about

the strictness of the rules as much as about the matrix. The P-onset error against experts moved from + 14 to + 1 ms with the reference. Choices inside the measurement chain can change values without any change in the patient (section 4.2). In PTB-XL pairs up to 24 h apart the point estimates of repeat variation were equal or lower than found here, with no primary class decided.

4.2. Repeat variation

Haq *et al* (2021) reported sinus-area-angle limits of -14.8° to $+14.5^\circ$ in same-day repeats. Our area-angle RC was 82.41° , but also 81.46° at identical recorded time and 46.64° in report-normal pairs, so elapsed time alone does not explain the spread; acquisition, segmentation, origin choice, signal quality and residual physiology can all contribute. Segmentation is not invariant to uniform gain (S0a): scaling all leads by 1.05 or 1.10 changed the area angle by 12.6° and 16.6° (root mean square, 385 ECGs; larger factors in S0a). Gain variation within the clinical pairs was not measured. Machine-reported measurements of the same recordings varied about as much for QRS duration and QT and less for QTcF (section 3.2); substantial repeat variation was also observed there, and these comparisons do not identify a shared variance component or its causes. The data do not separate measurement noise from state change, and a repeat within one recording would be the next comparator. The P-duration RC exceeded the between-person SD (47.84 versus 20.63 ms) without an ICC gain; no general conclusion about atrial physiology follows.

4.3. Matrix and boundaries

Ranking agreement and interchangeability are distinct: a Spearman correlation of 0.909 coexisted with 13.4% label switches. Zero stable outputs does not establish universal failure. For boundaries, the choice of reference changed the error estimates by amounts comparable to the errors themselves (P-onset median + 14 versus + 1 ms), favorably at some boundaries (P onset) and unfavorably at others (T onset); we did not compare reference effects with algorithm effects quantitatively. Against the Frank reference, inverse Dower and Kors differed in bias (section 3.5).

4.4. Other software and public controls

Comparisons of QRS and QTc classification across ECG programs, including LoopLens and BRAVEHEART, are reported in a related manuscript in preparation. The repeatability comparison of section 3.2 (S14d) is the only comparison of the two tools reported here; it is unfavorable to LoopLens in the markers where the intervals excluded 1 and implies no ranking of accuracy (the author developed LoopLens, and PTB-XL was used in its development). Directional recovery of QTcF in the randomized recordings does not validate other markers, as the route differences show.

4.5. What a clinical reader should take from this

The repeat variation, the matrix classes and the boundary errors describe LoopLens 10.1 on routine-care recordings; they are not treatment thresholds or minimal detectable changes. The clinical step still needed is a study of independent software on the same recordings, with blinded boundary marking including missing P waves and with repeat pairs in documented stable clinical state.

Limitations

This retrospective audit uses one clinical database from one centre with report proxies, informative missingness, complete-pair selection and enriched technical sampling, one implementation developed by the author, and devices whose effects are unresolved; the results do not generalize to other software, devices or populations. No comparator separates tool-related variation from clinical change: machine-reported measurements of the same recordings were compared post hoc (S3h), but a repeat within one recording was not performed, so the repeat variation cannot be apportioned between implementation and patients. Catalog shares count related identifiers as separate units (S3d). Clinical repeat differences combine state change with acquisition and processing error, and ICC depends on between-person variance in a non-random repeat sample. The proxy is selection-coupled. The frozen LoopLens release and the analysis scripts are public (Code availability), whereas further development of LoopLens remains private; aggregate results are in the Supplementary Material; historical generator identities remain incompletely documented (Table 1, S10a). ECG date-times come from each machine's internal clock, which in most cases was not synchronized with an external time source (Gow *et al* 2023), so the 0–2 h interval of a pair and its assignment to an admission and setting carry unquantified timing error.

The boundary comparison rests on LUDB, whose amplitudes are normalized per lead (no rescaling sensitivity was performed), which was used in development, with references that do not equal production fusion; the author's visual P-boundary screen was not independent and needs a blinded second marking that includes P waves missing from the available cases.

Native outputs are tied to the delineator's partly Kors-assisted fiducials, and matrix classes are technical rules on a fixed enriched sample; the replication criteria were prespecified on the historical classification including the 29 circular markers and were not recomputed without them. No result of this report is confirmed on data untouched by development: LUDB, PTB-XL and both internal development partitions were used or inspected during development, and the test partition had been opened for other research before the protocol was fixed, so the replication is internal and question-specific. PTB-XL repeats differ in interval, devices and chain, and the QTcF class sits on a band edge. No diagnostic, prognostic or treatment recommendation follows.

Conclusions

In one implementation, the catalog median ICC between recordings 0–2 h apart was moderate and varied widely between markers; more than half of the comparable non-circular catalog outputs met our technical criteria for matrix dependence and none met the stability criteria; and boundary errors depended on the expert reference. In PTB-XL pairs up to 24 h apart the point estimates of repeat variation were equal or lower than in the internal pairs, with no primary class decided. Repeat variation, reconstruction dependence and reference dependence should be assessed separately before physiological interpretation.

Declarations

Data availability statement

MIMIC-IV-ECG (version 1.0; Gow *et al* 2023; doi:10.13026/4nqg-sb35; ODbL 1.0), the PTB Diagnostic ECG Database (version 1.0.0; Bousseljot *et al* 1995, 2004; doi:10.13026/C28C71; ODC-By 1.0), PTB-XL (Wagner *et al* 2020, 2022; doi:10.13026/kfzx-aw45; CC BY 4.0), LUDB (version 1.0.1; Kalyakulina *et al* 2020, 2021; doi:10.13026/eegm-h675; ODC-By 1.0) and Autonomic Aging (version 1.0.0; Schumann and Bär 2021, 2022; doi:10.13026/2hsy-t491; ODbL 1.0) are openly available on PhysioNet (Pollard *et al* 2026). MIMIC-IV version 3.1 (Johnson *et al* 2023, 2024; doi:10.13026/kpb9-mt58; hosp and icu modules) is available to credentialed users under the PhysioNet Credentialed Health Data License 1.5.0. Derived patient-level data cannot be shared under the data use agreement; aggregate results are provided in the Supplementary Material. No individual signals, identifiers or private joins are redistributed. The test-partition replication is described in section 2.1 and section 3.7.

ECGRDVQ version 1.0.0 (ECGRDVQ 2016; doi:10.13026/C2HP45; Johannesen *et al* 2014; Pollard *et al* 2026) is openly available through PhysioNet under the Open Data Commons Attribution License v1.0. ECGCIPA version 1.0.0 (ECGCIPA 2018; doi:10.13026/C2967M; Vicente *et al* 2019; Pollard *et al* 2026), under the same license, was used only to audit waveform units. Aggregate public-control results are in Table 6 and Supplement S6b. Supplementary Data S2 comprises supplementary_data_S2_marker_reliability_public_v8.csv, supplementary_data_S2_marker_matrix_public_v8.csv, supplementary_data_S2_dictionary_v11.csv and supplementary_data_S2_marker_descriptions.csv, joined by unchanged marker identifiers; these files contain aggregate marker results only. A table linking each numerical statement to its aggregate source is available from the author.

Code availability

Analysis scripts and the frozen LoopLens release used for this study (PolyForm Noncommercial 1.0.0) are available at <https://github.com/leoprobst/ecg-auswertung-code> (release v1.0.0), in line with clause 8 of the PhysioNet Credentialed Health Data License 1.5.0; aggregate results are provided in the Supplementary Material. Further development of LoopLens remains private. The calculation state of each main analysis is in Table 1 and S10a.

Declarations

Ethical Approval and Consent to participate

This is a secondary analysis of de-identified research databases. The Institutional Review Boards of Beth Israel Deaconess Medical Center and the Massachusetts Institute of Technology approved the MIMIC-IV-ECG resource and waived individual consent (Gow *et al* 2023). The Institutional Review Board of Beth Israel Deaconess Medical Center approved the sharing of the MIMIC-IV research resource and granted a waiver of informed consent (Johnson *et al* 2023). Access to the credentialed MIMIC-IV modules was

obtained under the PhysioNet Credentialed Health Data Use Agreement after completion of the required CITI training. MIMIC-IV-ECG, the PTB Diagnostic ECG Database, PTB-XL, LUDB and Autonomic Aging are open-access PhysioNet resources, de-identified or anonymized by their providers before release; they were analyzed as anonymous data, and no attempt was made to re-identify individuals. LUDB reports written informed consent and approval by the Lobachevsky University IRB (#23; 19 October 2017; Kalyakulina *et al* 2021). The original ECGRDVQ and CiPA studies reported ethics approval and written informed consent (Vicente *et al* 2015, 2019). The present analysis recruited no participants and collected no new data. Formal ethics review of this secondary analysis was not sought because it used only de-identified or anonymized data that are public or accessed under their data use agreements.

Consent for publication

Not applicable. The manuscript contains no individual patient data, images or identifiers.

Availability of data and materials

MIMIC-IV-ECG (version 1.0; Gow *et al* 2023; <https://doi.org/10.13026/4nqg-sb35>; ODbL 1.0), the PTB Diagnostic ECG Database (version 1.0.0; Bousseljot *et al* 1995, 2004; <https://doi.org/10.13026/C28C71>; ODC-By 1.0), PTB-XL (version 1.0.3; Wagner *et al* 2020, 2022; <https://doi.org/10.13026/kfzx-aw45>; CC BY 4.0), LUDB (version 1.0.1; Kalyakulina *et al* 2020, 2021; <https://doi.org/10.13026/eegm-h675>; ODC-By 1.0), Autonomic Aging (version 1.0.0; Schumann and Bär 2021, 2022; <https://doi.org/10.13026/2hsy-t491>; ODbL 1.0), ECGRDVQ (version 1.0.0; ECGRDVQ 2016; <https://doi.org/10.13026/C2HP45>; Johannesen *et al* 2014) and ECGCiPA (version 1.0.0; ECGCiPA 2018; <https://doi.org/10.13026/C2967M>; Vicente *et al* 2019; used only to audit waveform units), the last two under the Open Data Commons Attribution License v1.0, are openly available on PhysioNet (Pollard *et al* 2026). MIMIC-IV version 3.1 (Johnson *et al* 2023, 2024; <https://doi.org/10.13026/kpb9-mt58>; hosp and icu modules) is available to credentialed users under the PhysioNet Credentialed Health Data License 1.5.0. Derived patient-level data cannot be shared under the data use agreement. Aggregate results are provided in the Supplementary Material and in four Supplementary Data S2 files (supplementary_data_S2_marker_reliability_public_v8.csv, supplementary_data_S2_marker_matrix_public_v8.csv, supplementary_data_S2_dictionary_v11.csv and supplementary_data_S2_marker_descriptions.csv), which contain aggregate marker results only. No individual signals, identifiers or private joins are redistributed. A table linking each numerical statement to its aggregate source is available from the author. Analysis scripts and the frozen LoopLens release used for this study (PolyForm Noncommercial 1.0.0) are available at <https://github.com/leoprobst/ecg-auswertung-code> (release v1.0.0), in line with clause 8 of the PhysioNet Credentialed Health Data License 1.5.0; aggregate results are provided in the Supplementary Material. Further development of LoopLens remains private.

Competing interests

The author is the developer of LoopLens, the software used for the analyses (non-financial interest). The author has no financial interest in LoopLens or in any related company or product, and no other competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The work was self-funded by the author.

Authors' contributions

L.P. conceived the study, developed the software, curated the data, performed the analyses, prepared the figures and wrote the manuscript. L.P. approved the final version.

Acknowledgements

The author thanks the PhysioNet team and the investigators who created and maintain MIMIC-IV, MIMIC-IV-ECG and the other open databases used.

Use of generative AI

The author used generative AI tools (Anthropic Claude and OpenAI Codex) to assist with analysis code, literature searches, data analysis workflows, drafting and editing of text, and figure preparation. AI-assisted output was reviewed, corrected where necessary and approved by the author, who takes full responsibility for the content. Patient-level analyses were run locally. AI tools are not listed as authors.

Registration and related work

The study was not prospectively preregistered. Dated retrospective plans and hashed specifications are not prospective registration. The physiological component map is incorporated into this audit and is not planned as a separate manuscript. An earlier version was submitted to medRxiv in October 2026 and was not posted (outside the server's scope); it has not been posted or published elsewhere. A related manuscript in preparation compares QRS and QTc classification across ECG programs, including LoopLens and BRAVEHEART; further analyses of the same database, including of the test partition, are also in preparation; none had been submitted or posted when this version was prepared, and no duplicated results are intended.

Glossary

ICC: summary of relative variance and agreement. RC: SD-based repeatability coefficient, 1.96 SD of paired differences including clinical variation (a descriptive spread, not a measurement-error estimate). D95: 95th percentile of the shortest angular distance between repeat values of a circular marker (a quantile, not comparable to RC). LoA: bias $\pm$ 1.96 SD, dispersion limits rather than confidence limits. P95: 95th percentile of absolute difference. Area angle: angle between integrated vectors. Peak angle: angle

at each loop's magnitude maximum. Minor-area class: rule based on opposite signed-area fractions. Net rotation: signed origin-closed area direction. Magnitude curvature: polarity-invariant quadratic descriptor. Conditional: intermediate technical class without clinical acceptance. Circular marker: angle on a circle, compared by shortest angular distance.

References

1. Andršová I, Hnatkova K, Toman O, Šišáková M, Smetana P, Huster KM, Barthel P, Novotný T, Schmidt G, Malik M 2022 Intra-subject stability of different expressions of spatial QRS-T angle and their relationship to heart rate. *Front Physiol* 13 939633 (10.3389/fphys.2022.939633)
2. Benchimol EI, Smeeth L, Guttman A, Harron K, Moher D, Petersen I, Sørensen HT, von Elm E, Langan S, M (RECORD Working Committee) 2015 The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med* 12 e1001885 (10.1371/journal.pmed.1001885)
3. Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement *Lancet* 327 307 – 10. 10.1016/S0140-6736(86)90837-8
4. Bousseljot R, Kreiseler D, Schnabel A (1995) Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. *Biomed Tech* 40(1):317–318. 10.1515/bmte.1995.40.s1.317)
5. Bousseljot R-D, Kreiseler D, Schnabel A 2004 PTB Diagnostic ECG Database (version 1.0.0) *PhysioNet* (10.13026/C28C71)
6. De la Garza, Salazar F, Egenriether B (2024) Exploring vectorcardiography: an extensive vectocardiogram analysis across age, sex, BMI, and cardiac conditions. *J Electrocardiol* 82:100–112. 10.1016/j.jelectrocard.2023.12.004)
7. ECGCIPA 2018 CiPA ECG validation study (version 1.0.0) *PhysioNet* (doi: 10.13026/C2967M)
8. ECGRDVQ (2016) ECG effects of ranolazine, dofetilide, verapamil, and quinidine (version 1.0.0) *PhysioNet* (10.13026/C2HP45)
9. Edenbrandt L, Pahlm O (1988) Vectorcardiogram synthesized from a 12-lead ECG: superiority of the inverse Dower matrix. *J Electrocardiol* 21:361–367. 10.1016/0022-0736(88)90113-6)
10. Gow B et al 2023 MIMIC-IV-ECG: diagnostic electrocardiogram matched subset (version 1.0) *PhysioNet* (10.13026/4nqg-sb35)
11. Haq KT, Lutz KJ, Peters KK, Craig NE, Mitchell E, Desai AK, Stencel NWL, Soliman EZ, Lima JAC, Tereshchenko L G 2021 Reproducibility of global electrical heterogeneity measurements on 12-lead ECG: the Multi-Ethnic Study of Atherosclerosis. *J Electrocardiol* 69 96–104 (10.1016/j.jelectrocard.2021.09.014)
12. Jekova I, Krasteva V, Leber R, Schmid R, Twerenbold R, Müller C, Reichlin T, Abächerli R 2016 Intersubject variability and intrasubject reproducibility of 12-lead ECG metrics: implications for human verification. *J Electrocardiol* 49 784–789 (10.1016/j.jelectrocard.2016.07.021)

13. Johannesen L et al (2014) Differentiating drug-induced multichannel block on the electrocardiogram: randomized study of dofetilide, quinidine, ranolazine, and verapamil. *Clin Pharmacol Ther* 96:549–558. 10.1038/clpt.2014.155)
14. Johnson AEW et al 2023 MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data* 10 1 (10.1038/s41597-022-01899-x)
15. Johnson A, Bulgarelli L, Pollard T, Gow B, Moody B, Horng S, Celi LA, Mark R 2024 MIMIC-IV (version 3.1) *PhysioNet* (10.13026/kpb9-mt58)
16. Kalyakulina AI, Yusipov II, Moskalenko VA, Nikolskiy AV, Kosonogov KA, Osipov GV, Zolotykh N, Yu, Ivanchenko MV 2020 LUDB: a new open-access validation tool for electrocardiogram delineation algorithms. *IEEE Access* 8 186181–186190 (10.1109/ACCESS.2020.3029211)
17. Kalyakulina A, Yusipov I, Moskalenko V, Nikolskiy A, Kosonogov K, Zolotykh N, Ivanchenko M 2021 Lobachevsky University Electrocardiography Database (version 1.0.1) *PhysioNet* (10.13026/eegm-h675)
18. Koo TK, Li MY (2016) A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 15:155–163. 10.1016/j.jcm.2016.02.012)
19. Kors JA, van Herpen G, Sittig AC, van Bommel JH (1990) Reconstruction of the Frank vectorcardiogram from standard electrocardiographic leads: diagnostic comparison of different methods. *Eur Heart J* 11:1083–1092. 10.1093/oxfordjournals.eurheartj.a059647)
20. Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, Roberts C, Shoukri M, Streiner DL (2011) Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *J Clin Epidemiol* 64:96–106. 10.1016/j.jclinepi.2010.03.002)
21. Makowski D, Pham T, Lau ZJ, Brammer JC, Lespinasse F, Pham H, Schölzel C, Chen SHA (2021) NeuroKit2: a Python toolbox for neurophysiological signal processing. *Behav Res Methods* 53:1689–1696. 10.3758/s13428-020-01516-y)
22. McGraw KO, Wong SP (1996) Forming inferences about some intraclass correlation coefficients. *Psychol Methods* 1:30–46. 10.1037/1082-989X.1.1.30)
23. Perez-Alday EA, Hamilton C, Li-Pershing A, Monroy-Trujillo JM, Estrella M, Sozio SM, Jaar B, Parekh R and Tereshchenko L 2018 The reproducibility of global electrical heterogeneity ECG measurements. *Comput Cardiol* 45 (10.22489/cinc.2018.162)
24. Platonov PG (2026) Rethinking automated P-wave analysis: a perspective on methodological limitations and future directions. *J Electrocardiol* 98 154415. 10.1016/j.jelectrocard.2026.154415)
25. Pollard T, Moody BE, Lehman L-WH, Gow BJ, Fernandes C, Xie C, Johnson A Mark R G and Heldt T 2026 PhysioNet as a global platform for biomedical research. *Nat Health* 1 792–795 (10.1038/s44360-026-00096-z)
26. Schreurs CA, Algra AM, Man S-C, Cannegieter SC, van der Wall EE, Schalij MJ, Kors JA, Swenne CA (2010) The spatial QRS-T angle in the Frank vectorcardiogram: accuracy of estimates derived from the 12-lead electrocardiogram. *J Electrocardiol* 43:294–301. 10.1016/j.jelectrocard.2010.03.009)

27. Schumann A, Bär K-J (2021) Autonomic aging: a dataset to quantify changes of cardiovascular autonomic function during healthy aging (version 1.0.0) *PhysioNet*. 10.13026/2hsy-t491
28. Schumann A, Bär K-J (2022) Autonomic aging: a dataset to quantify changes of cardiovascular autonomic function during healthy aging. *Sci Data* 9:95. 10.1038/s41597-022-01202-y)
29. Stabenau HF, Waks JW (2023) BRAVEHEART: open-source software for automated electrocardiographic and vectorcardiographic analysis. *Comput Methods Programs Biomed* 242 107798. 10.1016/j.cmpb.2023.107798)
30. Vicente J, Johannesen L, Mason JW, Crumb WJ, Pueyo E, Stockbridge N, Strauss DG (2015) Comprehensive T wave morphology assessment in a randomized clinical study of dofetilide, quinidine, ranolazine, and verapamil. *J Am Heart Assoc* 4:e001615. 10.1161/JAHA.114.001615)
31. Vicente J et al (2019) Assessment of multi-ion channel block in a phase I randomized study design: results of the CiPA phase I ECG biomarker validation study. *Clin Pharmacol Ther* 105:943–953. 10.1002/cpt.1303)
32. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative) (2007 The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies *Lancet* 370 1453-7 (10.1016/S0140-6736(07)61602-X)
33. Wagner P, Strodthoff N, Bousseljot R-D, Kreiseler D, Lunze FI, Samek W, Schaeffter T 2020 PTB-XL, a large publicly available electrocardiography dataset. *Sci Data* 7 154 (10.1038/s41597-020-0495-6)
34. Wagner P, Strodthoff N, Bousseljot R, Samek W, Schaeffter T 2022 PTB-XL, a large publicly available electrocardiography dataset (version 1.0.3) *PhysioNet* (10.13026/kfzx-aw45)
35. Young WJ, van Duijvenboden S, Ramírez J, Jones A, Tinker A, Munroe PB, Lambiase P, D, Orini M 2021 A method to minimise the impact of ECG marker inaccuracies on the spatial QRS-T angle: evaluation on 1,512 manually annotated ECGs. *Biomed Signal Process Control* 64 102305 (10.1016/j.bspc.2020.102305)

Figures

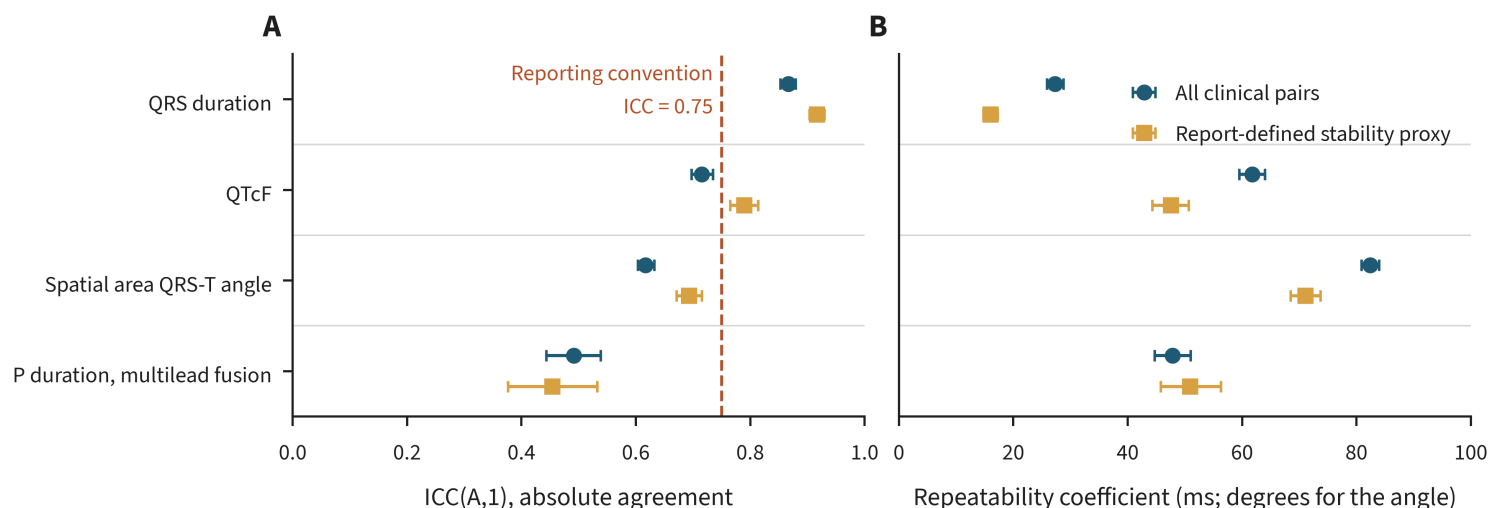


Figure 1

Short-interval reproducibility of four key measures in clinical ECG pairs 0-2 h apart. (A) ICC(A,1). (B) Repeat variation RC = 1.96 SD of the centered paired differences, in the unit of each measure (ms; degrees for the area QRS-T angle). Circles: all clinical pairs; squares: report-defined stability proxy. Bars: pointwise 95% percentile intervals from 1,000 person-level bootstrap draws (seed 77010). The dashed line marks the ICC ≥ 0.75 reporting convention, not a clinical certification. RC includes true clinical change as well as measurement error.

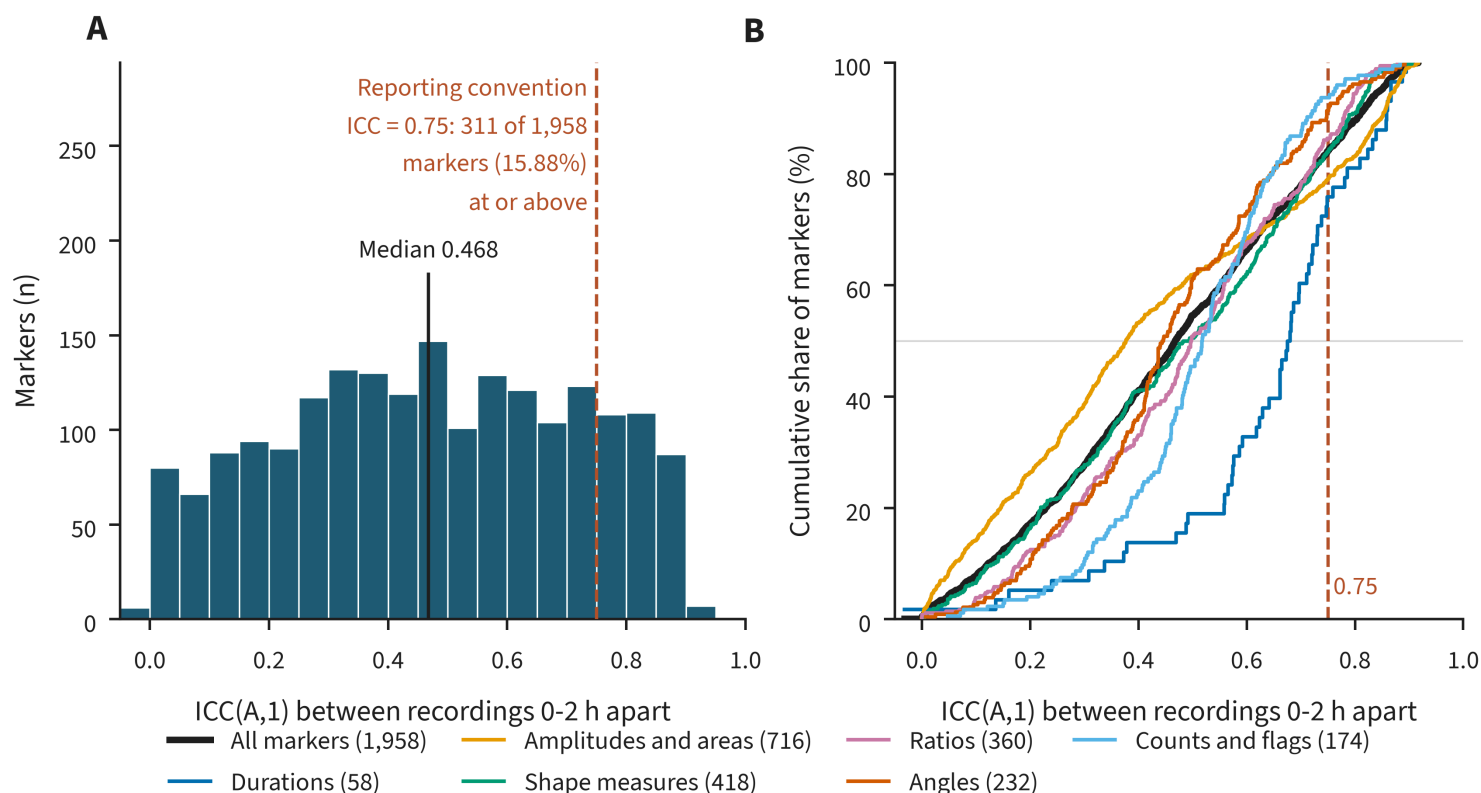


Figure 2

Distribution of ICC(A,1) across the 1,958 estimable non-circular markers between recordings 0-2 h apart (15,693 pairs, 10,187 persons). (A) Histogram with classes of 0.05. The solid line marks the median (0.468); the dashed line marks the ICC ≥ 0.75 reporting convention, which 311 markers (15.88%) reached. (B) Cumulative distribution for all markers (black) and for each marker family (colors; number of markers in parentheses). Each marker contributes one point estimate; single-marker intervals are in Supplementary Data S2. Six markers have a negative ICC. The 29 circular markers are excluded (section 2.3); the angle family has 232 markers here and 261 in Table S3d-1, which includes them. The ICC includes true clinical change and depends on the between-person spread of each marker; the 0.75 line is an author-chosen convention, not a clinical threshold.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ProbstECGVCGmeasurementauditsupplement.pdf](#)
- [FigS1catalogmatrixclasses.png](#)
- [FigS2matrixcontext.png](#)
- [FigS3LUDBboundaryerrors.png](#)
- [FigS4repairsdeltaauc.png](#)
- [supplementarydataS2dictionaryv11.csv](#)
- [supplementarydataS2markerdescriptions.csv](#)
- [supplementarydataS2markermatrixpublicv8.csv](#)
- [supplementarydataS2markerreliabilitypublicv8.csv](#)