WHITE PAPER

An AI-based framework to identify duplicate ECGs in clinical trials

The “ECG Mirror Participant” tool

Alain Gay, MD
Jean-Philippe Couderc, MBA, PhD
Luc Dekie, PhD
Todd Rudo, MD
Vickas Patel, MD, PhD
Hui Zhao, PhD

June 2026

CLARIO.

Contents

Introduction 3
The “ECG Mirror Participant” tool 4
Development of the “ECG Mirror Participant” tool 6
Real-world experience 7
Conclusions and perspectives 8
References 9

Introduction

In 2021, the U.S. Department of Justice reported a case in which a clinical study coordinator pleaded guilty to falsifying data across multiple drug trials, including the fabrication of participant enrollment records and clinical assessment measurements.[1] This type of misconduct not only violates the law but also undermines the scientific foundation of drug development, distorts safety and efficacy evaluations and erodes public trust in clinical research. Although deliberate fraud is rare, even isolated cases highlight a systemic vulnerability in the current system. Sponsors typically rely on data submitted by geographically dispersed clinical trial sites, but they have limited real-time visibility into how those data are generated.

In all clinical trials, electrocardiographic (ECG) signals must be collected reliably and accurately to properly assess safety of the investigational product. Specifically, ECG falsification in drug evaluation trials affects multiple stakeholders across the clinical research ecosystem. For example, participants are placed at risk when inaccurate ECG data obscure true cardiac safety signals or falsely suggest harm. These types of errors may lead to inappropriate dosing decisions, delayed access to effective therapies or unrecognized proarrhythmic risk.[2,3]

Inconsistencies in collection practices, protocol deviations or fabricated measurements may not be immediately apparent using conventional monitoring processes because of the large amount of data collected during these trials.

To address this gap, Clario developed an artificial intelligence (AI)-based analytical framework designed to identify anomalous patterns in ECG data submitted by clinical trial sites. Our approach applies convolutional neural networks and other machine learning tools to quickly identify "unexpected ECG data patterns." By analyzing ECG morphology patterns and variability across and within participants' ECG records, our system can flag unexpected recordings suggestive of substandard acquisition practices or potential data manipulation. This paper presents the scientific rationale, methodological foundation and implementation strategy of this unique AI-driven solution. Our objective is to provide sponsors with a proactive tool to enhance trial oversight, strengthen data integrity and reduce regulatory and ethical risk in clinical trials evaluating drug safety.

Over the past decade, AI-based facial recognition systems have demonstrated how machine learning can extract stable, person-specific features from highly variable real-world data. Across airports, smartphones, border control systems and social media platforms, these technologies reliably identify individuals despite differences in lighting, camera quality, facial expression, aging, partial occlusion or background noise. Rather than relying on superficial visual similarity, modern AI models learn deep, multidimensional feature representations that uniquely characterize each face. Once encoded, these features allow the system to cluster images belonging to the same person, even when acquired at different times or under different conditions. These algorithms can reliably detect mismatches, flag images that do not belong to a claimed identity and identify anomalies, such as spoofing attempts or corrupted data.

This paradigm of learning invariant identity-specific signatures from noisy biological signals extends well beyond facial recognition. The same conceptual framework can be applied to voice recognition,[4] fingerprint analysis,[5] gait detection[6] and body-surface electrocardiograms.[7] In this project, we elevated this AI-based identity modeling approach to the analysis of 12-lead ECG data collected in clinical trials, where the algorithm learns the intrinsic electrophysiologic "signature" of each study participant.

The "ECG Mirror Participant" tool

A "mirror participant" refers to a single individual being enrolled as multiple study participants in a clinical trial. In the following paragraphs, we present an illustrative example using intentionally altered (synthetic) data for educational purposes to demonstrate how Clario identifies mirror participants in a clinical trial.

To simulate mirror patients, we used data from a real clinical trial and intentionally exchanged the labels of three ECG strips between two study participants (ID 1 and 42) to mimic ECG mislabeling.

An example of the report output from the simulated data is shown in Table 1. The table provides a summary of the results for the overall set of ECGs included in the study and lists the ECG ID numbers for both the reference and comparison recordings, as well as the suspected mirror participants. We defined three levels of data integrity **(anomalous, suspect and normal)** based on thresholds of the dissimilarity score generated by the model between compared ECGs. A low score means that the compared tracings are considered to be very similar. In this specific example, the dataset includes one mirror participant (i.e., two different participant ID numbers with highly similar ECG morphologies) identified in the report table on the row classified as having an **"anomalous level of data integrity."** This row also reports the dissimilarity score threshold used to detect mirror participants. Although this threshold is applied consistently across same study types, it may be adjusted for other types, such as those involving long-term follow-up ECG recordings, where physiological changes within a participant may occur and impact the score.

Table 1. Mirror participant report for simulated data of a single clinical trial in which one ECG strip was mislabeled (switching the subject ID).

Data integrity levels	Thresholds dissimilarity score	Dataset occurrence (%)	ECGs (n=1290)	ECG ID	Mirror participents Source <> Mirror
Anomalous	<0.1	0.2%	3	ECG id-09 ECG id-11 ECG id-12	Subject #42 <> Subject #1
Suspect	0.1-0.5	0.3%	4	ECG id-78 ECG id-56 ECG id-85 ECG id-59	Subject #43 <> Subject 2 Subject #13 <> Subject 8 Subject #26 <> Subject 30 Subject #9 <> Subject 8
Normal	>0.5	99.5%	1283		

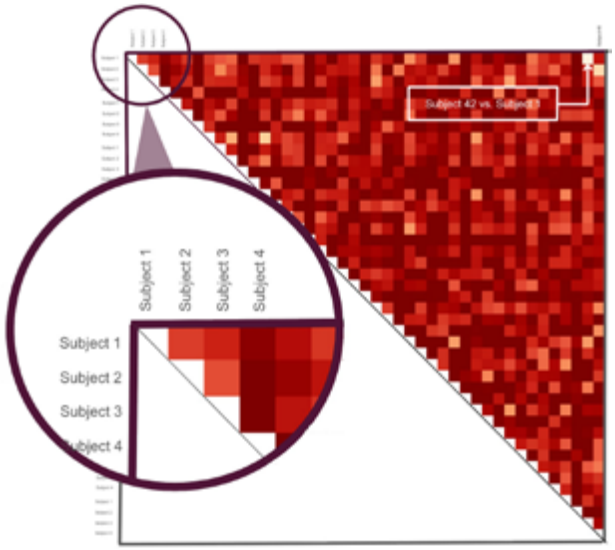
First, the AI model is used to quantify a **dissimilarity score** between ECG strips considering all pairs of ECG strips within and across study participants. A heatmap is then generated for data at the study level (Figure 1, Panel A) where a color scale is used to represent the values of the dissimilarity scores (low values in white and high values in dark red) in cells. If this step reveals suspicious data, then a second level of maps at the study participant level is generated (Figure 1, Panel B).

The study-level map (Figure 1, Panel A) provides a summary of the analysis of all possible ECG pairings generated across study participants. Each cell represents the minimum dissimilarity score obtained when comparing all ECG pairs between two participants. Under normal conditions, these minimum values are expected to remain high, reflecting clear differences between participants. Conversely, a low minimum dissimilarity score suggests unexpected similarity between ECGs from different participants, thereby raising concerns about the integrity of the ECG data (e.g., potential mislabeling or duplication). We call these occurrences "mirror participants" (as reported in Table 1).

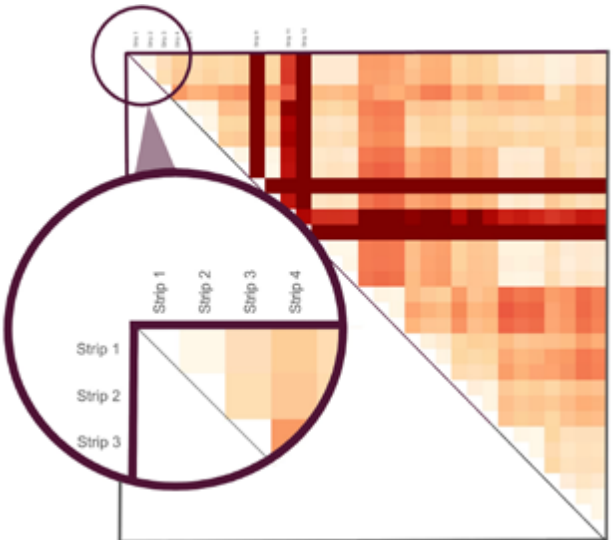
Accordingly, both axes represent participant ID numbers, and diagonal values equal 0, reflecting perfect dissimilarity scores as the two same ECGs are being compared. In a study-level map, one expects the minimum distance between different participants' ECG strips to be wide (dark red). In Panel A, we show the map for mislabeled ECGs between subject 1 and subject 42; the white "cell" shows the presence of ECG(s) with extremely low dissimilarity scores, namely, similar ECGs. This illustrates how the proposed tool is designed to visually identify potential ECG mislabeling among all ECG recordings collected for the study.

Figure 1: Examples of values of dissimilarity scores using two levels of analyses: **study-level map** (n=43 participants; Panel A) and **intra-participant map** (subject 42, n=30 ECGs; Panel B).

Panel A



Panel B



CLARIO.

The next step is to examine the ECG-specific dissimilarity scores for the participant of interest. In Figure 1 (Panel B), we present an **intra-participant map** illustrating the distribution of scores across all possible ECG pairings for a single participant (subject 42). In this map, the X and Y axes are ECG strip ID numbers. All cells are expected to appear in light colors, reflecting high similarity between the ECGs (from the same individual), unless significant changes in ECG morphology have occurred between ECGs. However, this example clearly shows the presence of three dissimilar ECGs within the same participant: strips 9, 11 and 12 (dark red horizontal and vertical lines in the map). These longitudinal fingerprints suggest that the associated ECG strips may not have been collected from the same participant since they are consistently associated with high dissimilarity scores when compared with all other ECGs collected from this participant and should be reviewed by a cardiologist.

Development of the "ECG Mirror Participant" tool

The mirror-participant tool was developed on a large, anonymized, aggregated set of ECGs collected over the past decades of Clario's support of clinical trials. The learning dataset consisted of ECGs extracted from 2,063 Holter recordings forming a pool of 308,697 vectorcardiography (VCG) signals from 621 individuals. The validation dataset consisted of 16,696 VCGs from 120 individuals, for whom there are 150 Holter recordings. Clario is currently applying the tool upon sponsor request to investigate if a specific study includes potentially anomalous data or when there is sponsor suspicion of improper site management.

Real-world experience

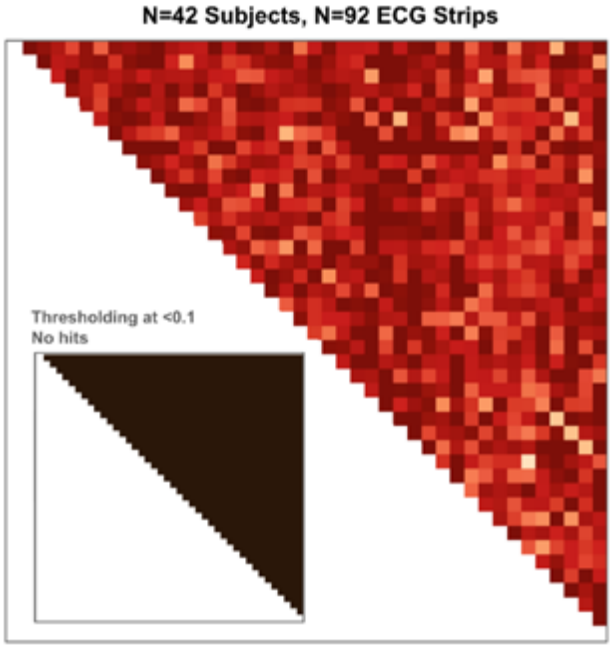
Two examples of real-world study-level maps from a Clario-partnered sponsor are presented in Figure 2. The maps shown were computed using ECGs collected from the same clinical study but from two different enrollment centers, Sites A and B (Panel A and Panel B, respectively). The study was a Phase III, randomized, double-blind, placebo-controlled, parallel-group trial involving patient populations rather than healthy volunteers. Panel A provides an example of a "clean" dataset with no evidence of ECG mislabeling. The map shows no light-colored cells, indicating that similarity scores across all participants fall within the expected range for this type of study.

In contrast, Panel B presents the map for the second site and reveals multiple light-colored cells, marked on the map using white arrows. These are regions of potentially suspicious or anomalous data. To better visualize the participants associated with these abnormal values, the map was filtered by removing all values ≥0.1. The resulting map is shown in the lower left corner of each panel. The site data shown in Panel B contains eight cells with extremely low dissimilarity scores, providing a clear and efficient way to identify potential mirror participants in a study. The filtered map in Panel A does not reveal any suspicious data.

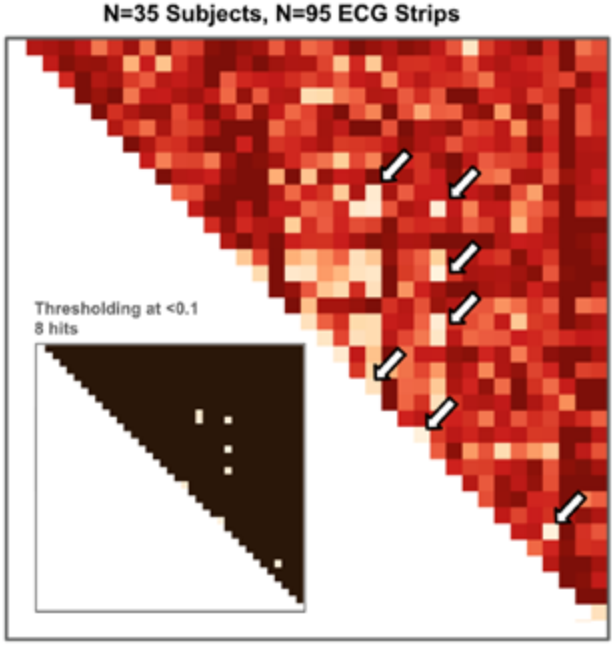
Once mirror-participants are suspected in the **study-level map**, our cardiologists review the associated **intra-participant maps** for potentially mislabeled individual ECG tracing pairs and review these ECGs individually to confirm whether they were collected from different participants. Using this tool, Clario can provide information about the presence or absence of highly similar ECGs to the sponsor, and at the sponsor's discretion, further investigation can be conducted.

Figure 2: Real-world examples of study-level maps from a Phase III study.

Panel A: Site A



Panel B: Site B



CLARIO.

Conclusions and perspectives

The proposed tool is unique and delivers an innovative method to strengthen the review of ECG data collected in clinical trials. The power of AI technologies is leveraged to review more tracings than one, or multiple cardiologist(s) could realistically process manually without excessive time commitment. Leveraging AI in this innovative way provides higher consistency and rigor than manual review while maintaining time efficiency, so large sets of ECGs can be reviewed. In addition, this tool addresses a real concern around the quality and integrity of clinical trial data, enhancing confidence in the data analyzed by sponsors and submitted to regulatory authorities.

The perspectives offered by this new method are extensive. One can readily envision its application in quantifying multiple dimensions of data quality at the enrollment-site level. It enables the establishment of benchmarks for the quality of ECG collection across different study types. The approach could be leveraged to assess the evolution of ECG morphologic changes over time as well as to characterize changes associated with drug exposure within large-scale datasets.

References

1. "Study Coordinator Pleads Guilty in Scheme to Falsify Clinical Drug Trial Data." U.S. Department of Justice, 26 October 2021, https://www.justice.gov/archives/opa/pr/study-coordinator-pleads-guilty-scheme-falsify-clinical-drug-trial-data (Accessed 28 May 2026).
2. "The Clinical Evaluation of QT/QTc Interval Prolongation and Proarrhythmic Potential for Nonantiarrhythmic Drugs." International Conference on Harmonisation, 12 May 2005, https://database.ich.org/sites/default/files/E14_Guideline.pdf (Accessed 28 May 2026).
3. Darpo, Borje. "The thorough QT/QTc study 4 years after the implementation of the ICH E14 guidance." *British journal of pharmacology* vol. 159,1 (2010): 49-57. doi:10.1111/j.1476-5381.2009.00487.x
4. Snyder, David, et al. 2018. "X-Vectors: Robust DNN Embeddings for Speaker Recognition." In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE Press, (2018): 5329–5333. doi:10.1109/ICASSP.2018.8461375
5. Cao, Kai, and Anil K Jain. "Automated Latent Fingerprint Recognition." *IEEE transactions on pattern analysis and machine intelligence* vol. 41,4 (2019): 788-800. doi:10.1109/TPAMI.2018.2818162
6. Chao, Hanqing, et al. "GaitSet: Regarding Gait As a Set for Cross-View Gait Recognition." *Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 33, no. 01, July 2019, pp. 8126-33, doi:10.1609/aaai.v33i01.33018126
7. Odinaka, Ikenna, et al. " ECG Biometric Recognition: A Comparative Analysis." *IEEE transactions on information forensics and security* vol. 7,6 (2012): 1812-1824. https://cse.buffalo.edu/~wenyaoxu/courses/fall2017/Paper/ECG_Biometrics2.pdf

About Clario

Clario is a leading provider of endpoint data solutions to the clinical trials industry, generating high-quality clinical evidence for life sciences companies. We offer comprehensive evidence-generation solutions that combine medical imaging, eCOA, precision motion, cardiac and respiratory endpoints.

For more than 50 years, Clario has delivered deep scientific expertise and broad endpoint technologies to help transform lives around the world. Our endpoint data solutions have been deployed over 30,000 times to support clinical trials in more than 100 countries. Our global team of science, technology, and operational experts have supported over 70% of all FDA drug approvals since 2015.