WHITE PAPER

An artificial intelligence–powered quality score tool for assessing ECG data

Luc Dekie, Ph.D.
Hui Zhao, Ph.D.
Robert Pastor
Jean-Philippe Couderc, Ph.D.
Robert Kleiman, M.D.
Vickas Patel, M.D., Ph.D.

November 2023

CLARIO.

Contents

Introduction

The probability of success (POS) for developing a new drug is 10% to 15%[1, 2] but varies with the therapeutic area. For example, the POS is as low as 3% in oncology and up to 33% in ophthalmology.[1] The main causes for drug failure are lack of clinical efficacy (40%-50%), unmanageable toxicity (30%), poor drug-like properties (10%-15%) and poor strategic planning or lack of commercial need (10%).[2] Most new drug entities, or approximately 70%, fail in Phase I or the first-in-human (FIH) trial predominantly due to unmanageable toxicity or poor pharmacokinetics.[2] These findings highlight the importance of the FIH trial in drug development. In an FIH trial, the major causes for failure are assessed and dictate if an asset progresses to late-phase development or the whole program fails.

Given the critical position an early-phase trial has in determining the POS of a new drug entity, it is essential that the data generated at this phase are complete, usable and reliable. Any steps taken to ensure data quality in the early phase will increase the return on investment if it prevents an asset from failing or clearly supports terminating a program.

In addition, if manageable toxicity is detected in early-phase development, the steps taken to mitigate the toxicity in later-phase development can only be successful if the early-phase data are complete and reliable. For these reasons, assessing data quality after completion of the first dose cohort may help to identify errors in data collection or documentation. This provides an opportunity to determine the cause of the errors and implement corrective actions before more impactful data are collected from higher dose cohorts, during which toxicity is more likely to arise.

In clinical medicine, an electrocardiogram (ECG) provides more information about patient safety than any other single assessment. Assuring the ECG data are usable and complete in early-phase development allows for reliable assessment of the effect of a new compound on the QT interval and other ECG parameters. This improves the probability of successfully obtaining a waiver to conduct a thorough QT (TQT) study that can reduce the time and expense to complete the drug development program.[3] If the strategy selected is to collect, clean and store the ECG data for later analysis (e.g., once the pharmacokinetics of the drug have been assessed and the drug candidate passes proof of concept), then quality checking the full dataset considerably improves the likelihood the future analysis will provide meaningful results.[4]

Reviewing and assessing the quality of extensive continuous ECG data can be a laborious task. Nonetheless, both the current availability of powerful computerized tools, such as artificial intelligence (AI) and machine learning (ML), and the existence of extensive ECG repositories have made this challenge more attainable. In addition, the standardized signal-based nature of the ECG lends itself well to AI/ML modeling and automation for handling large datasets.

The marriage of AI/ML and ECGs has ushered in a new era of cardiac healthcare, promising precision, efficiency and the potential for improving trial outcomes. This application is gaining increasing relevance, especially with the growing adoption of wearable devices such as ECG patches and other ambulatory devices, making it an area of considerable importance for cardiac monitoring.

In more recent years, AI/ML has also become a natural extension of the classical signal processing paradigm, in which the linear processing blocks are replaced by non-linear equivalents that enable scientists to handle a much broader set of problems.[5] One area of interest is the ability to quickly respond to the collection of ECG tracings and signals in case the quality is unacceptable or to provide a benchmark of ECG quality for large sets of data.

These algorithms are typically based on deep learning (DL) network architectures comprising multiple hidden layers.[6] The most frequently used DL algorithms are convolutional neural networks (CNNs), which were originally proposed for object recognition and image classification. However, AI/ML tools can also be useful to rapidly assess data quality and support signal processing tasks.

Clario has developed an AI-powered tool to automatically check the quality of long continuous ECG recordings without the need for manual review of the recordings. Using an AI-powered tool will increase the overall quality of ECG analysis, as well as reduce the risk of errors due to missing data or inappropriately misclassifying useful data.

AI-powered ECG quality score

ECG dataset

To create an automated process to assess ECG quality, we leveraged an AI model to identify a 3.6-second ECG segment, referred to hereafter as an "epoch," during which signal stability and clarity would yield accurate interval duration measurements. The model was trained on an anonymized dataset comprised of epoch examples, each labeled as either good or poor quality. Epochs were labeled "good" if the beats included in the epoch were measurable using the 3 beats on lead II (3BL2) methodology,[7] and epochs were labeled "poor" if the beats included were unable to be measured due to inadequate quality. The example epochs were divided into two independent sets, one set for training and validation and another set for testing (Figure 1). The testing set consisted of independent studies from the training and validation sets.

Figure 1. Training, test and validation dataset

Training		Validation		Test	
Data source		**Data source**		**Data source**	
# Epochs	84,502	**# Epochs**	11,925	**# Epochs**	257,648
# Studies	80	**# Studies**	74	**# Studies**	38
# Participants	929	**# Participants**	753	**# Participants**	1,210

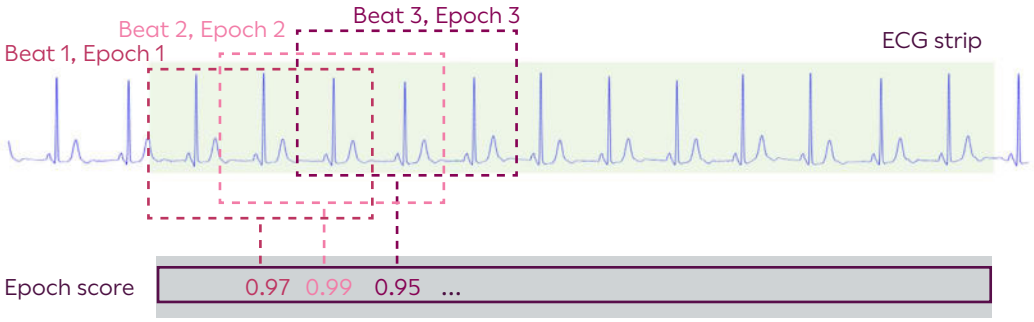
AI-powered epoch quality score

The DL model is designed based on the ResNet architecture.[8] It consists of three main parts:

1. Input: convolutional layer and batch normalization layer
2. Residual: three residual units, each composed of two convolutional layers with batch normalization, activation functions and optional dropout
3. Output: a fully connected layer with the sigmoid activation function to produce the final quality epoch score

This DL model takes a 3.6-second ECG epoch centered around a specific beat as input and outputs the quality score for that epoch (Figure 2). Scores range from 0 to 1, where the quality is the highest at 1.

Figure 2. Illustration of the AI-powered quality score method



The continuous ECG signal is decomposed in a series of 3.6-second ECG epochs centered on each cardiac beat. The AI-powered tool delivers a quality score for each epoch.

The quality score can also be used to locate sections on a Holter recording containing high-quality data (Figure 3).

Figure 3. ECG signal and AI-powered quality scores

The continuous ECG recording is depicted in the lower panel, alongside a time-synchronized mapping of the epoch quality scores, visualized using a color scale (upper panel). In this mapping, the color orange represents lower epoch scores, while green is used to denote the highest epoch scores.

Performance

To evaluate the model's performance, we used an anonymized, independent test dataset containing 257,648 epochs. With the test dataset, the model classified 215,075 of the epochs as "good" and 33,840 of the epochs as "poor" (Figure 4). We then compared the AI-powered model assessments to those made using the standard 3BL2 methodology. From these comparisons, we calculated four performance metrics defined as follows:

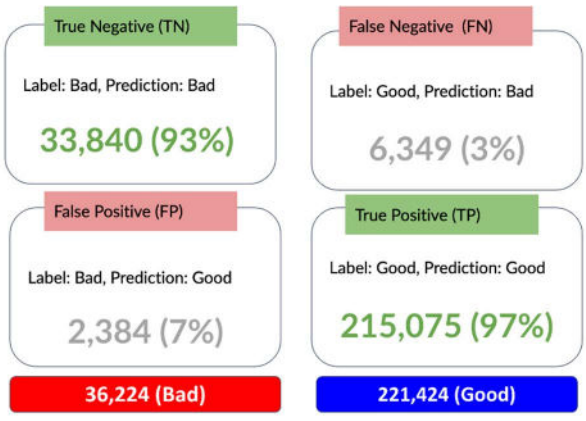
1. True positive (TP): The beats contained in a good epoch detected by the AI-powered model were found to be measurable using the 3BL2 methodology.

2. False positive (FP): The AI-powered model indicated an epoch to be of good quality, but the beats included were not measurable using the 3BL2 methodology.

3. True negative (TN): The beats contained in a poor epoch detected by the AI-powered model were found not to be measurable using the 3BL2 methodology.

4. False negative (FN): The AI-powered method indicated an epoch to be of poor quality, but the beats included were measurable using the 3BL2 methodology.

Figure 4: Epoch quality score testing results



These results suggest that the model has a sensitivity of 97%, meaning it accurately identified good-quality epochs. Additionally, it exhibited a specificity of 93%, indicating its capability to correctly classify poor-quality epochs. The FP and FN values were 7% and 3%, respectively.

From epoch quality to ECG quality assessment

In collect, clean and store studies, which do not undergo immediate analysis by cardiac safety technicians, we are interested in assessing whether ECG strips would be usable once they are extracted from the Holter recordings. The epoch quality score provides a solution by predicting, with high precision, when a measurable 3-beat sequence exists. To ensure a strip contains at least one measurable epoch, we consider the full strip to be usable if it contains at least one epoch with a quality score greater than the critical usability score that was determined by fitting samples from a validation dataset to achieve the desired sensitivity and specificity.

Initial experience with the AI-powered quality score tool

To assess how the AI-powered tool is performing in the real-world environment, we used anonymized datasets from three randomly selected healthy-volunteer Phase I studies from the Clario EXPERT database that were conducted at three different Phase I units. Two studies were single ascending dose (SAD) studies, and the third was a multiple ascending dose (MAD) trial. These are the two trial designs we see most often in collect, clean and store studies. In the SAD studies, a 24-26–hour continuous Holter recording was recorded on Day 1 of each cohort. SAD Study 1 consisted of 58 participants, and 15 time points were planned to be extracted per participant. SAD Study 2 included 56 participants with a maximum of 10 time points planned per participant. In the MAD study, a 24-26–hour continuous Holter recording was recorded on Day 1 and at steady state. The study included 56 participants with a maximum of eight time points planned to be extracted per participant.

We compared the ability of the new AI method to identify at least 3 ECG strips per time point with that of our traditional methodology used in hundreds of SAD, MAD and TQT studies with healthy volunteers.[4] Our traditional method uses an advanced computer-assisted, statistical process to extract ECGs from continuous recordings. During the per-protocol–specified ECG extraction windows, up to 10 non-overlapping digital 12-lead ECG tracings are extracted from continuous recordings. If the method is not able to extract ECGs, the entire extraction window is reviewed manually to identify good-quality beats.

From these comparisons, we calculated the TP, FP, TN and FN, which were defined as follows:

- True positive (TP): Both the AI-powered model and traditional method identified ≥3 good quality replicates per time point.
- False positive (FP): The AI-powered model indicated ≥3 replicates could be extracted, but the traditional method generated <3 replicates.
- True negative (TN): Both the AI-powered model and traditional method indicated <3 replicates could be extracted per time point.
- False negative (FN): The AI-powered method indicated <3 replicates could be extracted, but the traditional method reported ≥3 replicates could be extracted.

As seen in Table 1, the AI-powered model had 100% specificity for detecting time points that do not generate sufficient usable replicates. Importantly, the AI model also did not produce false positive results by labeling time points as usable that would in truth be unusable for analysis. The true positive rate varied between 97.8% and 99.8%, and the false negative rate varied between 0.2% and 2.2%.

Table 1: Summary of the TN, FN, FP and TP values from the AI model used on three studies

	N	True negative (TN)	False negative (FN)	False positive (FP)	True positive (TP)
SAD study 1	58	4 (100%)	19 (2.2%)	0 (0%)	847 (97.8%)
SAD study 2	56	1 (100%)	1 (0.2%)	0 (0%)	558 (99.8%)
MAD study	56	2 (100%)	18 (1.2%)	0 (0%)	1,535 (98.8%)

Discussion

During the past decade, the convergence of machine learning and ECGs represents a paradigm shift in how cardiac disease is assessed. The rich history of AI is now contributing to this transformative partnership, offering a glimpse into a future where cardiac pathophysiology is detected earlier, treated more effectively and ultimately results in less morbidity and mortality through the power of technology and medical science.

Every drug in development will have to undergo testing to evaluate its potential to prolong the QT interval. Early-phase ascending dose studies offer a great opportunity to collect ECG data to investigate the effect on QT interval prolongation. In these studies, high drug exposures are obtained that potentially cannot be reached in later studies. As most new drugs fail in Phase I, sponsors may decide not to analyze the ECG data until they have a better understanding of the pharmacokinetics and efficacy of their compound. If the strategy selected is to collect, clean and store the continuous ECG data for later analysis, it is important to assure the data are of good quality and will be useful to support a potential TQT waiver. In contrast, ECG quality assessed during ongoing studies after completion of the first dose cohorts will also help identify any errors and allow time for corrective actions (e.g., site retraining, device replacement) to be put in place when the higher dose cohorts are assessed, when the potential for QT interval prolongation and toxicity is higher.

In our paper, we present an application of AI/ML to assess the quality of continuous ECG tracings collected in early-phase clinical trials. Clario is leveraging this new technology in early-phase studies to assess data quality in near real time. Using this approach, we demonstrated that the AI-powered model always correctly identified time points with poor quality (true negatives=100%) without the risk of polluting the overall dataset by labeling time points of too low quality as usable time points (false positives=0%). Furthermore, the false negative rate was also low (0.2%-2.2%). Performing a manual review of this small subset of data in a collect, clean and store study will provide assurance that the collected data will be usable for future analysis, if required.

Clario is actively integrating and developing AI/ML technologies, with a focus on enhancing its services in the management, review and analysis of ECG data. This effort aims to ultimately improve data processing speed and data accuracy as well as to assist in the interpretation of data, promising exciting future applications in this domain.

References

1. Wong CH, Siah KW, Lo AW. Estimation of Clinical Trial Success Rates and Related Parameters. Biostatistics 2019;20(2): 273–86. https://doi.org/10.1093/biostatistics/kxx069.
2. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical Development Success Rates for Investigational Drugs. Nature Biotechnology 2014;32(1): 40–51. https://doi.org/10.1038/nbt.2786.
3. ICH E14. "Questions & Answers (R3)," December 10, 2015. https://database.ich.org/sites/default/files/E14_Q%26As_R3_Q%26As.pdf.
4. Darpo B, Borin M, Ferber G, Galluppi GR, Hopkins SC, Landry I, et al. ECG Evaluation as Part of the Clinical Pharmacology Strategy in the Development of New Drugs: A Review of Current Practices and Opportunities Based on Five Case Studies. The Journal of Clinical Pharmacology 2022;62(12): 1480–500. https://doi.org/10.1002/jcph.2095.
5. Askin S, Burkhalter D, Calado G, El Dakrouni S. Artificial Intelligence Applied to Clinical Trials: Opportunities and Challenges. Health and Technology 2023;13(2): 203–13. https://doi.org/10.1007/s12553-023-00738-2.
6. Mathew A, Amudha P, Sivakumari S. Deep Learning Techniques: An Overview. In Advanced Machine Learning Technologies and Applications, edited by Hassanien AE, Bhatnagar R, Darwish A, 1141:599–608. Singapore: Springer Singapore, 2021.
7. Badilini F, Sarapa N. Implications of Methodological Differences in Digital Electrocardiogram Interval Measurement. Journal of Electrocardiology 2006;39(4): S152–6. https://doi.org/10.1016/j.jelectrocard.2006.05.030.
8. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–8. Las Vegas, NV: IEEE, 2016. https://doi.org/10.1109/CVPR.2016.90.

About Clario

Clario is a leading healthcare research and technology company that generates the richest clinical evidence in the industry for our pharmaceutical, biotech and medical device partners. Across decentralized, hybrid and site-based trials, our deep scientific expertise, global scale and the broadest endpoint technology platform in the industry allows our partners to transform lives. Clario has the only technology platform that combines eCOA, cardiac safety, medical imaging, precision motion, and respiratory endpoints. Clario's global team of science, technology and operational experts have helped deliver over 24,000 trials and contributed to over 500 FDA and EMEA new drug approvals involving more than five million patients in 120 countries. Our innovation has been transforming clinical trials for 50 years.