Healthcare Transformation with Informatics and Artificial Intelligence
J. Mantas et al. (Eds.)

doi:10.3233/SHTI230534

Data Quality in Healthcare for the Purpose of Artificial Intelligence: A Case Study on ECG Digitalization

Arian RANJBAR[a,1] and Jesper RAVN[a]
[a] *Medical Technology and E-health, Akershus University Hospital, Norway*

Abstract. The quantity of data generated within healthcare is increasing exponentially. Following this development, the interest of using data driven methodologies such as machine learning is on a steady rise. However, the quality of the data also needs to be considered, since information generated for human interpretation may not be optimal for quantitative computer-based analysis. This work investigates dimensions of data quality for the purpose of artificial intelligence applications in healthcare. Particularly, ECG is studied which traditionally rely on analog prints for initial examination. A digitalization process for ECG is implemented, together with a machine learning model for heart failure prediction, to quantitatively compare results based on data quality. The digital time series data provide a significant accuracy increase, compared to scans of analog plots.

Keywords. Data Quality, Digitalization, Artificial Intelligence, Machine Learning

1. Introduction

The amount of generated data worldwide is exponentially increasing every year, including within healthcare. With the rapid development of data driven methodologies, there has recently been an increased interest in using healthcare data to develop new analysis methods, diagnostics and treatments. Although methods such as deep learning have shown capabilities of interpreting noisy data, good overall performance still relies on quality data, obeying the principle of "garbage in, garbage out", [1]. Additionally, data quality is often task-dependent, and what amounts to useful data for human interpretation, often dominant in healthcare settings; may not necessarily correlate with quantitative machine-reliant tasks.

Most healthcare data have been collected with traditional operation in mind, [2]. This is particular evident in e.g. analog ECG data, still prevalent among hospitals today. Analog ECGs provide ease-of-access for humans, with good enough resolution for initial examination. However, a scanned version may not provide detail enough for more advanced modelling, especially targeting beyond human performance.

This work aims to raise awareness regarding data quality within healthcare, for the purpose of artificial intelligence. Particularly by performing a case study delimited to

[1] Corresponding Author: Arian Ranjbar, Akershus Universitetssykehus HF, 1478 Lørenskog, Norway; E-mail: arian.ranjbar@ahus.no. This research was partially funded by Nasjonalforeningen for Folkehelsen. This work is part of a project with unanimous approval in ethical trial by Helsodirektoratet (21/39600), chair: Elisabeth Sagedal.

ECG data. First by presenting related work, then by performing quantitative experiments comparing models using digitalized data versus data originating from analog plots intended for human interpretation; on the task of heart failure predictions.

1.1. Related Work

Assessment of data quality is often seen as a multi-variate evaluation, along several properties, dependent on the end-use, [3]. Quality properties is often divided into categories affecting availability, usability, reliability, relevance and presentation quality, [4]. Although all aspects of quality needs to be considered, reliability is of extra importance when training machine learning models, [5]. Reliability can further be branched into accuracy, i.e. that the representation well reflects the true state and will not cause ambiguity; consistency, such that formats and value domains match; integrity, i.e. format is clear and meets specification; and completeness, whether the deficiency of a component affects data accuracy overall.

In supervised learning, accuracy not only regards the input data, but also the labels. For example, in the case of cardiovascular disease, previous studies have shown that patients get erroneously diagnosed using standardized diagnose codes, [6]. This bias will inherently be adopted into supervised learning models using inaccurate labels, e.g. by training a classifier on ECG data to detect the conditions. Similarly, completeness not only affects overall performance but may also affect accuracy on stratified subsets of the test data. For example, the training set may lack coverage on a characteristic such as gender or age, which may lead to biased results during inference, [7].

Objectively measuring quality is challenging since the underlying properties may be heavily task-dependent and lack quantitative measures. For example, completeness with regards to age can be controlled for; but quantitatively estimating coverage of the input domain in ECGs is not feasible, since the space is infinite and the distribution is not known beforehand. General metrics can be used for simple ratio measurements of specific properties, [3]; however, for aggregated evaluation, task-specific experiments are often necessary, [4].

2. Method

As a case for data quality in the context of artificial intelligence, an investigation into ECG data is made. As previously mentioned, ECG instruments typically use digital recording but presents output analogly as paper prints, before eventually being scanned for digital storage. Although machine learning has been successfully applied to such paper scans, [8], there is a significant information loss in the storage process amounting to lower data reliability. In addition, the process risk other compounding errors due to the manual interventions, e.g. error in manual entries.

To increase reliability, a digitalization process is implemented for ECG data, building upon two components. First, a scanner is installed to all ECG measuring systems. The scanner allows for identification through patient ID wristbands and automatic transfer of relevant information to the ECG system. Then an automatic transfer process is implemented, directly transferring the digital ECG to the storage system, rather than having the instrument print the ECG before being scanned for storage.

Evaluation of increase in reliability is done according to the following metrics. Resolution is used as a proxy for accuracy and completeness in the signal. Consistency,

and integrity in the signal is qualitatively evaluated based on the used formats. Additionally, the ratio of correct coupling between ECGs and patient ID is evaluated for label quality and consistency.

Although these metrics serve as an initial evaluation, it is challenging to predict their effect in aggregate for the purpose of artificial intelligence. Thereby, two comparable machine learning experiments are implemented, to estimate task specific performance based on data quality. Specifically, a multi-channel 1D and single channel 2D Resnet model of eight blocks, [9], is implemented for the digital time series and ECG plots respectively; performing heart failure prediction. The ECG plots are emulated using digital plots in the theoretically maximal resolution from the analog setup, estimating an upper bound on the performance. Generating the plots also ensure no performance boost is gained through variance in the underlying training data.

3. Results

The implementation was carried out at Akerhus University Hospital, where on average 300 ECGs are collected every day. The previous system, although using 1,000 Hz sampling, ploted 10 s ECGs with 300 dpi over 25 cm, which gives an upper bound of 20,000 dots. However, the scans are done with approximately 1,000 pixels in width, which significantly limits the resolution. The digital time series is stored in its original 1,000 Hz sampling, increasing the resolution by at least a factor 10. Furthermore, the prints may vary in format depending on instrument manufacturer and operator, e.g. by channels per page, plot scale and layout; whereas the time series in stored in a standardized format. In the previous system, up to 20% of incoming paper scans had the wrong identifiers (from manual entry), prevented in the new system using wristband scans.

3.1. Heart failure prediction

The experiment to evaluate task performance based on data quality was carried out through a dedicated research platform, [10]. 10,000 ECGs were collected, 5,000 from healthy patients and 5,000 from patients with heart failure. Machine learning models were implemented in Tensorflow 2.11, trained using Adam optimizer with learning rate 0.001 and batch size of 32. 20% of the data was used for testing and 15% of the training data was used for validation. The model trained on digitalized time series data achieved AUC 0.92, whereas the model trained on image data of ECG plots achieved AUC 0.85.

4. Discussion

Quality metrics of reliability improved after the digitalization implementation, and more notably it affected the machine learning experiments. However, the comparison has several limitations, as a real production setting would have significantly more tuned models. For example, channel extraction could be done on the ECG plots using traditional image processing before eventually applying time series-based architecture. Such modelling has been carried out successfully, [8]. On the other hand, this increases the engineering complexity to account for the lower data reliability, and beats the purpose

of this experiment. Similarly, training datasets and hyperparameters could be more carefully designed, to increase performance.

4.1. Other dimensions of quality

This study mainly focuses on reliability, however other dimensions of quality is also of importance in a healthcare setting. Availability regards accessibility and timeliness of information, and have proven to be non-trivialities. Healthcare systems often rely on IT silos, with separate storage of electronic health records and sensor data etc., using combinations of legacy and modern systems, [10]. Digitalization of the data to standardized formats allow access unconstrained by physical location and lowers latency in information broadcasting among the healthcare systems, essential for emergency care but also for timely decision support by algorithms.

5. Conclusion

The increasing quantity of data within healthcare enables use of data driven methods such as machine learning. However, the quality aspect is of equal importance and not as widely considered. This study investigated dimensions of data quality of interest for machine learning modelling within healthcare, particularly information reliability. To demonstrate the concept, a digitalization process was implemented for ECG data showing increase in data accuracy, consistency, integrity and completeness. In addition, two comparable machine learning models were implemented for the digitalized time series data and ploted image data respectively, showing higher performance in the former. Finally, other dimensions of quality were discussed, e.g. availability, which is challenging to provide in healthcare systems due to the underlying IT infrastructure.

References

[1] Kilkenny MF, Robinson KM. Data quality:“Garbage in–garbage out”. SAGE Publications Sage UK: London, England; 2018.
[2] El Khatib M, Hamidi S, Al Ameeri I, Al Zaabi H, Al Marqab R. Digital Disruption and Big Data in Healthcare-Opportunities and Challenges. ClinicoEconomics and Outcomes Research. 2022:563-74.
[3] Pipino LL, Lee YW, Wang RY. Data quality assessment.Communications of the ACM. 2002;45:211-8.
[4] Cai L, Zhu Y. The challenges of data quality and data quality assessment in the big data era. Data science journal. 2015;14.
[5] Gudivada V, Apon A, Ding J. Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. International Journal on Advances in Software. 2017;10:1-20.
[6] Ofstad A, Johansen O, Brunborg C, Morkedal B, Fagerland M, Laugsand L, et al. The validity of heart failure diagnoses at hospital-discharge and ambulatory evaluation visits: insights from two Norwegian local hospitals. European Heart Journal. 2021;42;724-851.
[7] Ranjbar A, Skolt K, Aakenes Vik KT, Sletvold Øistad B, Wermundsen Mork E, Ravn J. (in press). Fairness in Artificial Intelligence: Regulatory Sandbox Evaluation of Bias Prevention for ECG Classification. Medical Informatics Europe. 2023.
[8] Mishra S, Khatwani G, Patil R, Sapariya D, Shah V, Parmar D, et al. ECG paper record digitization and diagnosis using deep learning. Journal of medical and biological engineering. 2021;41(4):422-32.
[9] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770-8.
[10] Ranjbar A, Ravn J, Ronningen E, Hanseth O. (in press). Enabling Clinical Trials of Artificial Intelligence: Infrastructure for Heart Failure Predictions. Medical Informatics Europe. 2023.