Estimating the Minimal Size of Training Datasets Required for the Development of Linear ECG-Lead Transformations

Daniel Guldenring[1], Ali Rababah[2], Dewar D Finlay[2], Raymond R Bond[2], Alan Kennedy[2], Michael Jennings[2], Khaled Rjoob[2], James McLaughlin[2]

[1]HS Kempten, Kempten, Germany
[2]Ulster University, Belfast, United Kingdom

Abstract

Linear electrocardiographic lead transformations (LELTs) are used to estimate unrecorded ECG leads by applying a number of recorded leads to a LELT matrix. Such matrices are commonly developed using a training dataset. The size of the training dataset has an influence on the estimation performance of a LELT matrix. However, an estimate of the minimal size required for the development of LELTs has previously not been reported.

The aim of this research was to determine such an estimate. We generated LELT matrices from differently sized (from n = 10 to n = 540 subjects in steps of 10 subjects) training datasets. The LELT matrices and the 12-lead ECG data of a testing dataset (n = 186 subjects) were used for the estimation of Frank VCGs. Root-mean-squared-error values between recorded and estimated Frank leads of the testing dataset were used for the quantification of the estimation performance associated with a given size of the training dataset.

The performance of the LELTs was, after an initial phase of improvement, found to only marginally improve with additional increases in the size of the training dataset. Our findings suggest that the training dataset should have a minimal size of 170 subjects when developing LELTs that utilise the 12-lead ECG for the estimation of unrecorded ECG leads.

1. Introduction

Linear electrocardiographic (ECG) lead transformations (LELTs) are used to estimate or derive unrecorded target leads by applying a number of recorded basis leads to a LELT matrix [1, 2].

These LELTs are a well-established concept in computerized electrocardiology and are used in a wide range of different application areas. Application areas of LELTs include the estimation of the Frank vectorcardiogram (VCG) using standard 12-lead ECG [3] or Mason-Likar 12-lead ECG data [1]. A further application area of LELTs is their use in reduced lead systems that estimate the 12-lead ECG from a reduced number of monitoring compatible ECG leads [4]. An emerging application area of LELTs is the performance assessment of patch based wearable electrocardiographic devices [5].

The most common form of LELTs utilizes transformation matrices that are designed to be used on ECG data of the general adult population. Such transformation matrices are commonly developed using a training dataset that is composed of ECG data obtained from a number of different subjects. For each of these subjects, one set of target leads and basis leads is included in the training dataset. The transformation matrices of LELTs are typically developed using multivariate linear regression analysis on the target leads and basis leads of the training dataset [1, 3]. The number of subjects whose ECGs are included in the training dataset is commonly referred to as the size of the training dataset.

It is desirable that the transformation matrices of LELTs are capable of producing accurate estimates of the target leads for all members of the adult population. The utilization of unrepresentative and small training datasets is known to yield transformation matrices that perform poorly in the general adult population. Training datasets should therefore be of sufficient size in an attempt to accurately reflect the statistical relationship between the basis leads and the target leads of the general population.

Recording a large training dataset for the development of a new transformation matrix is potentially a time and cost expensive procedure. It would therefore be desirable to know an estimate of the minimal training dataset size that is required for the development of LELT matrices. However, such an estimate has, to the best of our knowledge, not previously been reported in the literature.

The aim of this research is to determine an estimate of the minimal training dataset size required for the development of LELTs. To this end, we assess the estimation performance of LELT matrices developed using training datasets of increasing size. We define the minimal required size of the training dataset as a size, at which only

ISSN: 2325-887X DOI: 10.22489/CinC.2021.342

marginally improvements in the performance of a LELT matrix can be achieved through further increases in the size of the training dataset.

2. Material and methods

2.1. Study population

We base our research on a study population of 726 subjects. The study population is composed of 229 normal subjects, 265 subjects with myocardial infarction and 232 subjects with left ventricular hypertrophy. The study population was randomly partitioned into a test dataset ($DTest$) of fixed size and a pool of 540 subjects ($DTrain$) that were used to assemble training datasets of varying size. Table 1 details the composition of $DTest$ and $DTrain$.

Table 1. Composition of the test data ($DTest$) and the train data ($DTrain$).

	Normal	MI	LVH	Total
$DTest$	59	66	61	186
$DTrain$	170	199	171	540

Notes. ***Normal****, Subjects with no abnormalities in their ECGs;* ***MI****, Subjects with myocardial infarction;* ***LVH****, Subjects with left ventricular hypertrophy.*

2.2. Target and basis leads of the LELTs

The eight independent leads I, II, V1 to V6 of the standard 12-lead ECG were chosen as the basis lead set for the LELT matrices that were assessed in this research. This was because the standard 12-lead ECG is the most widely adopted ECG recording format [6], which makes the standard 12-lead ECG a popular basis lead set that is used in different LELTs.

The heart-vector model [7] of the cardiac electrical activity provided the rational for the utilization of the Frank VCG as the target lead set for the LELT matrices assessed in this research. Any ECG lead can, in accordance with the heart-vector model, be expressed as a weighted sum of the orthogonal X, Y and Z leads used by the Frank VCG. The minimal size of the training dataset required for the development of a LELT matrix, that is used for the estimation of the Frank VCG, was therefore regarded as a good estimate of the minimal training size required for the development of any other LELT matrix.

2.3. BSPM data

One body surface potential map (BSPM) was recorded for each of the 726 subjects in the study population. Each BSPM used in this research contains electrocardiographic data of 120 BSPM leads. A representative average QRS-T complex was calculated for each of the 120 BSPM leads. Three of the 120 leads were recorded from electrodes placed on the right and left wrist and the left ankle (VR, VL and VF respectively). Electrodes situated at 81 anterior and 36 posterior locations were used to record 117 thoracic leads. All thoracic leads were recorded with reference to the Wilson central terminal (WCT). A comprehensive description of the BSPM data and the recording procedure can be found in [8]. A Laplacian 3D interpolation procedure [9] was applied to the 117 thoracic BSPM leads. This was performed to obtain body surface potentials at the locations of the 352 Dalhousie torso [10] nodes.

2.4. Frank VCG data

One Frank VCG [11] was extracted from each of the 726 BSPMs. First, body surface potentials at the A, C, E, F, H, I and M electrode locations of the Frank lead system were extracted from the interpolated BSPM data. Body surface potentials from body locations that were not a direct subset of the 352 Dalhousie torso nodes were obtained using linear interpolation [12]. Second, the body surface potentials at the Frank electrode locations were used to derive the Frank VCG using (1).

$$\boldsymbol{VCG} = [\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}] = [\boldsymbol{\varphi_A}, \dots, \boldsymbol{\varphi_M}] \cdot \boldsymbol{A^T}. \quad (1)$$

Where $\boldsymbol{\varphi_A}$, $\boldsymbol{\varphi_C}$, $\boldsymbol{\varphi_E}$, $\boldsymbol{\varphi_F}$, $\boldsymbol{\varphi_H}$, $\boldsymbol{\varphi_I}$, and $\boldsymbol{\varphi_M}$ are $n \times 1$ vectors that contain n sample values of potentials at the Frank electrode locations A to M respectively, $[\cdot]^{\boldsymbol{T}}$ refers to the transpose of a matrix, n denotes the number of samples in the average QRS-T complex, $\boldsymbol{A}$ is a 3×7 matrix of published coefficients [13] that allow for a derivation of the Frank VCG using the potentials $\boldsymbol{\varphi_A}$ to $\boldsymbol{\varphi_M}$, and $\boldsymbol{VCG}$ is a $n \times 3$ matrix containing n sample values of the Frank VCG, the $n \times 1$ vectors $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$ contain n sample values of the three Frank leads X, Y and Z respectively.

2.5. Standard 12-lead ECG data

One standard 12-lead ECG was extracted from each of the 726 BSPMs. First, body surface potentials recorded at the wrists and ankles were used to calculate the limb leads of the standard 12-lead ECG as well as the potential at the WCT. Second, the body surface potentials at the electrode locations associated with the precordial leads were extracted from the interpolated BSPM data. Required body surface potentials from locations that were not a direct subset of the 352 Dalhousie torso nodes were obtained using linear interpolation. Third, average QRS-T complexes of the precordial leads were calculated in reference to the WCT using body surface potentials obtained from the locations of the precordial electrodes.

2.6. Linear regression based ECG lead transformation matrices

The data in $DTrain$ was used to assemble training datasets of different sizes. More precisely, training datasets staring from n = 10 to n = 540 subjects were generated in steps of 10 subjects. Random sampling with replacement was used to compose 200 different instances of each training dataset size using the data in $DTrain$. The different training dataset instances were used to generate a total of 200 transformation matrices for each training dataset size. Transformation matrices that allow for the estimation of the Frank VCG from the standard 12-lead ECG were developed using the multivariate linear regression based approach in (2).

$$ {}_{m}\boldsymbol{AVCG_i} = ({}_{m}\boldsymbol{S12L_i^T} \cdot {}_{m}\boldsymbol{S12L_i})^{-1} \cdot {}_{m}\boldsymbol{S12L_i^T} \cdot {}_{m}\boldsymbol{VCG_i}. \tag{2} $$

Where $[\cdot]^{\boldsymbol{T}}$ and $[\cdot]^{\boldsymbol{-1}}$denote the transpose and the inverse of a matrix respectively, ${}_{m}\boldsymbol{AVCG_i}$ refers to a 8× 3 matrix of transformation coefficients that allows for the transformation of the eight independent leads I, II and V1 to V6 of the standard 12-lead ECG into the Frank VCG, $m \in \{10,\ldots,540\}$ denotes the size of the training dataset, n refers to the number of QRS-T sample values in the training dataset of size m, $i \in \{1,\ldots,200\}$ denotes the instance of the training dataset that was used for the development of ${}_{m}\boldsymbol{AVCG_i}$, ${}_{m}\boldsymbol{VCG_i}$ refers to a n× 3 matrix that contains n sample values of the X, Y and Z leads of the Frank VCG and ${}_{m}\boldsymbol{S12L_i}$ refers to a n× 8 matrix that contains n sample values of the eight independent leads I, II and V1 to V6 of the standard 12-lead ECG.

2.7. Derivation of the target leads

The ${}_{m}\boldsymbol{AVCG_i}$ matrices were used to derive the target leads of the 186 subjects in $DTest$. This was performed using the approach in (3) and for all LELT matrices with $i \in \{1,\ldots,200\}$ and $m \in \{10,\ldots,540\}$.

$$ {}_{m}\boldsymbol{dVCG_i} = \boldsymbol{S12L} \cdot {}_{m}\boldsymbol{AVCG_i}. \tag{3} $$

Where ${}_{m}\boldsymbol{AVCG_i}$, m and i are as defined in (2), $\boldsymbol{S12L}$ is a $n \times 8$ matrix that contains the n sample values of the QRS-T complex for the eight independent leads of the standard 12-lead ECG of one subject in $DTest$ and ${}_{m}\boldsymbol{dVCG_i}$ is $n \times 3$ matrices that contain the derived leads of the Frank VCG.

2.8. Performance assessment

The average performance of each ${}_{m}\boldsymbol{AVCG_i}$ matrix was quantified using the data of the 186 subjects in $DTest$. First, root mean square error (RMSE) values were calculated between the QRS-T complexes of the recorded and the derived target leads. This was performed for each transformation matrix and for each of the 186 subjects in $DTest$. Second, the mean of the 186 different RMSE values associated with each target lead and ${}_{m}\boldsymbol{AVCG_i}$ matrix was determined for each $i \in \{1,\ldots,200\}$ and $m \in \{10,\ldots,540\}$. The outcome of this performance assessment was a 200×54 matrix of ${}^{RMSE}_{m}\overline{\boldsymbol{VCG}}_i$ elements. Where each ${}^{RMSE}_{m}\overline{\boldsymbol{VCG}}_i$ contains one mean RMSE value for each of the three Frank VCG leads, $i \in \{1,\ldots,200\}$ and $m \in \{10,\ldots,540\}$ respectively denote the instance and size of the training dataset hat was used for the development of the ${}_{m}\boldsymbol{AVCG_i}$ matrix associated with the mean RMSE values in ${}^{RMSE}_{m}\overline{\boldsymbol{VCG}}_i$.

2.9. Determination of the minimal size required for the training dataset

The minimal required size of the training dataset was determined separately for each Frank lead using two different criteria. Both criteria defined the minimal required size using mean ${}^{RMSE}_{m}\overline{\boldsymbol{VCG}}_i$ values that were calculated over all $i \in \{1,\ldots,200\}$ instances of a given training dataset size m.

The first criterion was based upon right-tailed t-tests (significance level alpha = 0.05) that were used to test the null hypothesis, that the mean ${}^{RMSE}_{m}\overline{\boldsymbol{VCG}}_i$ value associated with a given training dataset size m was equal or less than 101% of the mean ${}^{RMSE}_{540}\overline{\boldsymbol{VCG}}_i$ value. This test was performed for each training dataset size $m \in \{10,\ldots,540\}$. A failure to reject the null hypothesis corresponds to a lack of statistical evidence that the mean ${}^{RMSE}_{m}\overline{\boldsymbol{VCG}}_i$ value associated with a given training size m is at least +1% greater than the mean ${}^{RMSE}_{540}\overline{\boldsymbol{VCG}}_i$ value. The smallest size m at which a t-test failed to reject the null hypothesis was considered as the minimal required training dataset size m.

The second criterion for defining the minimal required size of the training dataset was based upon reaching 95% of the reduction in the mean ${}^{RMSE}_{m}\overline{\boldsymbol{VCG}}_i$ value that was observed between the smallest (m = 10 subjects) and the largest (m = 540 subjects) training dataset size. First, the difference between the mean ${}^{RMSE}_{10}\overline{\boldsymbol{VCG}}_i$ value and the mean ${}^{RMSE}_{540}\overline{\boldsymbol{VCG}}_i$ value was calculated. This difference was regarded as the maximal reduction in the mean ${}^{RMSE}_{m}\overline{\boldsymbol{VCG}}_i$ value that can be achieved when increasing the size of the training dataset from 10 to 540 subjects. Second, the differences between the mean ${}^{RMSE}_{m}\overline{\boldsymbol{VCG}}_i$ values for $m \in \{10,\ldots,540\}$ and the mean ${}^{RMSE}_{540}\overline{\boldsymbol{VCG}}_i$ value were calculated. Third, these differences were expressed as percentage of the maximal reduction in the mean ${}^{RMSE}_{m}\overline{\boldsymbol{VCG}}_i$ value. Fourth, a right-tailed t-test (significance level alpha = 0.05) was used to test the null hypothesis, that the remaining reduction in the mean ${}^{RMSE}_{m}\overline{\boldsymbol{VCG}}_i$ value was equal or less than 5 % of the maximal observed reduction. This test was performed for each training dataset size $m \in \{10,\ldots,540\}$. A failure to reject the null hypothesis corresponds to a lack of statistical evidence that remaining reduction in the mean ${}^{RMSE}_{m}\overline{\boldsymbol{VCG}}_i$ value was greater than 5 % of the maximal value. The smallest size m for which this test was not able to reject the null hypothesis was considered as the minimal required training dataset size m.

3. Results

A summary of the findings from the analysis of the minimal required size of the training dataset is provided in Table 1.

Table 1. Minimal required training dataset sizes for each Frank VCG lead and mean ${}^{RMSE}_{m}\overline{\boldsymbol{VCG}}_{i}$ values associate with the minimal required and the maximal training dataset size.

derived lead	criterion[a]	min. size	mean (min. size)[b]	mean (size 540)[c]
X	1% of final value	170	30.4	30.0
	95 % reduction	170	30.4	
Y	1% of final value	120	30.8	30.3
	95 % reduction	130	30.7	
Z	1% of final value	130	47.6	46.9
	95 % reduction	130	47.6	

[a]criterion used for the determination of the minimal required training dataset size; [b]mean ${}^{RMSE}_{m}\overline{\boldsymbol{VCG}}_{i}$ value in µV associated with the minimal required training dataset size; [c]mean ${}^{RMSE}_{m}\overline{\boldsymbol{VCG}}_{i}$ value in µV for a training dataset size of 540 subjects.

4. Discussion and conclusion

This paper reported on the assessment of the minimal training dataset size that is required for the development of LELT matrices. Our analysis was conducted on LELT matrices that transform the standard 12-lead ECG into the Frank VCG. A minimal training dataset size of 170 subjects, 130 subjects and 130 subjects was found to be sufficient for the estimation of Frank leads X, Y and Z respectively. Any ECG lead can, in accordance with the heart-vector model [7], be expressed as a weighted sum of the orthogonal Frank X, Y and Z leads. We therefore conclude that a training dataset size of 170 subjects should be sufficient for the development of LELTs that utilize the 12-lead ECG for the estimation any ECG lead that can be recorded from the body surface.

A limitation of this research is that the assessed LELT matrices were developed and tested on ECG data that was obtained from three equally represented cohorts (normal subjects, subjects with myocardial infarction and subjects with left ventricular hypertrophy). Whether the presence of different additional cardiac disorders in the training and testing datasets would have an influence on the required minimal training dataset size has not been assessed in this research.

A further limitation of this research is that it has solely assessed the influence of the training dataset size on the mean estimation performance (mean ${}^{RMSE}_{m}\overline{\boldsymbol{VCG}}_{i}$ value). This is a limitation as the assessed LELT matrices were intended to be used with all members of the adult population. Such matrices should therefore not only have an acceptable mean estimation performance but should ideally also perform equally well for all members of the adult population. The subject-to-subject variability of the estimation performance of a LELT matrix is therefore an additional performance metric that has to be considered. Future research should thus investigate the influence of the training dataset size on this subject-to-subject variability.

References

[1] Guldenring D, Finlay DD, Strauss DG, et al. Transformation of the Mason-Likar 12-lead electrocardiogram to the Frank vectorcardiogram. Conf Proc IEEE Eng Med Biol Soc. 2012;2012:677-680.

[2] Guldenring D, Finlay DD, Bond RR, et al. The derivation of the spatial QRS-T angle and the spatial ventricular gradient using the Mason-Likar 12-lead electrocardiogram. J Electrocardiol. 2015;48(6):1045-1052.

[3] Kors A, van Herpen G, Sittig AC, et al. Reconstruction of the Frank vectorcardiogram from standard electrocardiographic leads: diagnostic comparison of different methods. Eur Heart J. 1990;11(12):1083–1092.

[4] Guldenring D, Finlay DD, Nugent CD, et al., Estimation accuracy of a reduced lead system during simulated ischemia. Computing in Cardiology. 2011. p. 237-240.

[5] Guldenring D, Finlay, DD Bond RR, et al. Which part of the P-QRS-T is best when developing linear ECG-lead transformations for the performance assessment of patch based electrocardiographic devices?, J Electrocardiol. 2019 57(Supplement):101.

[6] Ribeiro AH, Ribeiro MH, Paixão GMM, *et al.* Automatic diagnosis of the 12-lead ECG using a deep neural network. Nat Commun. 2020;11:1-9.

[7] Burger HC, van Milaan JB, Heart-vector and leads. Part II, Br Heart J. 1947:9(3):154-160.

[8] Montague TJ, Smith ER, Cameron DA, Rautaharju PM, Klassen GA, Felmington CS, et al. Isointegral analysis of body surface maps: surface distribution and temporal variability in normal subjects. Circulation 1981; 63(5): 1166-1172.

[9] Oostendorp TF, van Oosterom A, Huiskamp G. Interpolation on a triangulated 3D surface. J Comput Phys 1989; 80(2):331-343.

[10] Horáček BM. Numerical Model of an Inhomogeneous Human Torso. Adv Cardiol 1974; 10:51-57.

[11] Frank E. An accurate, clinically practical system for spatial vectorcardiography. Circulation 1956; 13(5):737-749.

[12] Schijvenaars BJA, Kors JA, van Herpen G, Kornreich F, van Bemmel JH. Interpolation of body surface potential maps. J Electrocardiol 1995; 28 Suppl 1: 104-109.

[13] Macfarlane PW. Lead systems. In: Macfarlane PW, van Oosterom A, Pahlm O, Kligfield P, Janse M, Camm J, editors. Comprehensive Electrocardiology. 2nd ed. United Kingdom. London: Springer; 2011; p. 375-426.

Address for correspondence:

Daniel Guldenring
Room T117, Faculty of Electrical Engineering, HS Kempten,
Bahnhofstraße 61, 87435 Kempten, Germany
daniel.gueldenring@hs-kempten.de