Automated Extraction of Large-Scale Data from Electronic Health Records in Tertiary Hospital

1Chaithra N, 2Madhu B, 3Manjunath Basavaraju, 4Balasubramanian S

1Assistant Professor, 2Associate Professor, 3Founder of Mysoftlabs, 4Director (Research) and Dean
Division Medical Statistics, Faculty of Life Sciences, JSS Academy of Higher Education & Research, Mysore, India

Abstract

Background: Cardiovascular diseases are the major cause of mortality globally, as well as in India. Large volumes of data are continuously generated from clinical and diagnostic studies catalogued in EHRs. The collection and storage of a large amount of unstructured and heterogeneous data from various sources are not easily proceeded as in simple data bases; therefore, we aimed to develop a fast and efficient algorithm to establish a research platform from the EHRs System for prediction of cardiovascular disease in South India.

Methods: A retrospective study was designed to access the echocardiography reports from EHRs, who underwent transthoracic echocardiography at JSS Hospital. The Macro VBA algorithm was applied to 16580 records of patients with a total of 55 attributes. During data cleansing, repetitive records, spelling errors and unrealistic data are detected and removed by using Macro VBA. The missing values in the data are identified and replaces them with the mean value of the data. Box and Whisker plot distribution of raw quantitative echocardiographic parameters, Multivariate identity, and extreme outliers are existing in the data series.

Results: The study presents the automated extraction of non-standardized transthoracic echocardiographic reports from EHRs by using the Macro VBA algorithm and it was applied to 6891 outpatients and 9689 inpatients records at a single academic medical center in the year 2016. The total time taken by Macro VBA for extracting 16580 echocardiography reports and generating output in a structured format was less than an hour on a personal laptop, this algorithm will serve to decrease the burden of obtaining large scale data in a structured format.

Conclusion: The outcomes presented in this paper are the first ever designed algorithm to linking large scale data from EHRs to transform the information from an unstructured into a structured form by using Macro VBA for investigating quality of care and improving cardiovascular diseases.

Index terms - Cardiovascular Disease, Electronic Health Records (EHRs), Echocardiographic data, Macro VBA.

I. INTRODUCTION

Cardiovascular diseases (CVDs) have become the leading cause of mortality in India. They are caused by disorders of the heart and blood vessels; and include coronary heart disease, cerebrovascular disease, high blood pressure, rheumatic heart disease and heart failure [1,2]. The Global Burden of Disease study estimates of age-standardized CVD death rate of 272 per 100 000 population in India is higher than the global average of 235 per 100 000 population, 52 % of CVD deaths occur before the age of 70 years. Indians current cardiovascular epidemic is increasing with rapid economic development and changing lifestyles [3].Cardiovascular diseases are one of the highest-flying diseases of the modern world [4,5]; the treatment of the said disease is quite high and not affordable by most of the patients particularly in India [6]. The diagnosis of this disease has several tests like Echocardiography test (ECHO), TreadMill Test (TMT), Electrocardiogram (ECG), Holter Monitoring (HM) test, Transoesophageal echocardiography (TEE). These tests can help doctors diagnose heart and blood vessel diseases and conditions in adults and children. Echocardiography (ECHO), is an ultrasound test used to view moving pictures of the heart on a screen. It is used to detect and evaluate a variety of conditions, including heart valve problems, abnormal heart rhythms, heart disease, heart murmurs or infections involving the heart. Echocardiography provides a continuous stream of huge volumes of biomedical data found in unstructured or non-standardized clinical narrative reports [7,8].

Large volumes of data are continuously generated from clinical notes and diagnostic studies catalogued in Electronic Health Records (EHRs) [9,10]. Data based on electronic health records are rich with individual level and are becoming an increasingly common data source for clinical risk prediction. Harnessing, information from EHRs for research requires a full understanding of data set their linkages, management and data quality in large data sets, which presents unique opportunities for analysis and challenges. The collection and storage of a large amount of unstructured and heterogeneous data from various sources are not easily proceeded as in simple data bases; therefore, we aimed to develop a fast and efficient data extraction tools and technique [11,12]. A Macro is a code written in VBA (Visual Basic for Applications) that allows running a chunk of code whenever it is executed, it helps technicians to easily incorporate user-written functions into a spreadsheet. This study presents a Macro VBA program capable of large-scale transformation of heterogeneous echocardiographic reports into a structured data format. The application of visual basic for extracting information from sequential data in biomedical sources has the potential to impact both clinical practice and research [13,14].
II. METHODS

1. Study area

JSS Hospital is one of the biggest hospitals in India and has one of the biggest critical and emergency care facilities with 260 beds. The hospital has state-of-the-art infrastructure and the most advanced equipment, located in an area of 12.5 acres and has a built-up space of 12.5 lakh sq ft. The hospital provides service in 37 specialties/super specialties and has 67 special clinics with 1,800 beds under one roof to fulfill the demand for healthcare services for needy people in and around the Mysore region. To cater to the ever-increasing health care needs of Mysore and the surrounding population, the hospital provides advanced and affordable healthcare to all sections of the society and it is a non-profit hospital dedicated to serving the poor and downtrodden with affordable and quality healthcare. This hospital was selected because of its outstanding general practice based primary care units and unique integrated electronic health information system. Particularly, the system captured rich information on cardiovascular disease and also other information from primary care units. This enables us to use comprehensive data for demographical, epidemiological and predicting cardiovascular disease in South India.

2. Data sources and study population

Electronic health record system includes data sources for demographic characteristics, clinical data, healthcare services, laboratory services, inpatients and outpatient information are recorded in backbone software. A retrospective study of 45,111 transthoracic echocardiography reports from the year 2014 to 2016 in a tertiary hospital. Echocardiography data are one of the large datasets that are available in the electronic health record database, it includes 29 qualitative and 26 quantitative measurements that were taken during the echocardiography examination. Fig. 1. represents the year wise distribution of the three tests i.e. ECHO, ECG and TMT which was performed between 2014 to 2016, these tests were done to find whether the patients were suffering from cardiovascular diseases.

![Fig.1. Distribution for ECHO, ECG and TMT tests performed at JSS Hospital from 2014 to 2016.](image)

Diagnosis of CVDs involves the application of certain tests like ECHO, ECG, and TMT based on the analysis of the investigations done from 2014 to 2016. This graph shows that a total of 11982, 15904, 17225 ECHO records, 4369, 4583, 3386 ECG and 2066, 2471, 3035 TMT reports are recoded in year 2014, 2015 and 2016 respectively. The growth in the use of ECHO increased dramatically in the hospital from 2014 to 2016 and occurred at a higher rate than ECG and TMT test for the evaluation of cardiovascular diseases.

3. Procedure for Echocardiographic examination

Echocardiographic examinations were performed by an experienced ECHO technologist using Philips HD11XE, CX30 echocardiography system, with an S8-3, S4-2 transducer. The flow chart of echocardiography test in diseases diagnosis are displayed in Fig. 2. Who visited the cardiac medical check-up unit in the hospital. Measurements on echocardiography were obtained according to the recommendation of the American Society of Echocardiography (ASE) guidelines and comprehensive transthoracic echocardiography examination guidelines by the Indian Academy of Echocardiography. All studies were done with patients lying in the left lateral decubitus position and breathing quietly. The transthoracic echocardiography report should be comprised of the following sections: (1) Demographic and other identifying information, (2) Echocardiographic (M-mode, Doppler, and Two-Dimensional Imaging) evaluation, and (3) Impressions.

3.1 Demographic and other identifying information: The echocardiography report includes the following demographic and other identifying information: (1) Patient’s IP number (2) Date and time on which the study was performed (3) Age (4) Gender (5) Location of the patient (6) Referring physician identification (7) Interpreting cardiologist identification (8) Name or identifying information for persons performing the study (ECHO technicians, Physician); (9) Echocardiographic instrument identification and other identifying information.

3.2 Echocardiographic (M-mode, Doppler and Two-Dimensional Imaging) evaluation: The following cardiac and vascular structures are generally evaluated as part of a comprehensive transthoracic echocardiography report by using M-mode, Doppler and Two-Dimensional Imaging : Aortic Root; Left Atrium; Right Ventricle; Left Ventricle Internal Diameter during Diastole/Systole; Intact Ventricular Septum Diameter during Diastole/Systole; Left Ventricular Posterior Wall Diameter during Diastole/Systole; End Diastolic/Systolic Volume; End Volume; Stroke Volume; Ejection Fraction(%); Fractional Short(%); Mitral valve; Mitral regurgitation; Tricuspid valve; Tricuspid.
regurgitation; Aortic Valve - The maximal aortic jet velocity; Aortic regurgitation; Pulmonary Vascular - The maximal Pulmonary jet velocity; Pulmonary regurgitation; Left Ventricle; Left Atrium; Right Ventricle; Right Atrium; Aorta; Pulmonary Artery; IVS; IAS; Mitral valve; Aortic valve; Tricuspid valve; Pulmonary valve; Pericardium; Colour Doppler; Doppler Study; Other. It should be emphasized that the identification and measurement of some of the structures listed may not always be possible or necessary to provide a comprehensive, clinically relevant report. However, it is important for the echocardiography cardiologist to comment on the M-mode, Doppler, and Two-Dimensional Imaging findings. In general, quantitative measurements are often performed and frequently adequate for interpretation.

3.3 Impressions based on Two-Dimensional Imaging: Impressions of the echocardiographic report are based on Two-Dimensional Imaging method and includes statements that: Answer the questions posed by the referring physician and emphasize abnormal findings of the echocardiography [7, 15]

![Flowchart of the echocardiography process](image)

**Fig. 2.** The process of echocardiography test in disease diagnosis.

4. **Data Analysis Plan**

The Cardiology Department historically has generated a large scale of data, driven by record keeping, compliance, regulatory requirements, and patient care. While earlier most data are stored in textual form and not usable for statistical evaluations, the current trend is toward the rapid digitization of these large volumes of data in EHRs to transform the information from an unstructured into a structured form. The challenge will be how to mine the vast datasets to facilitate and expedite decision making in healthcare. These are the steps in meeting this challenge which could be used to develop tools and techniques for automated analysis of functional datasets.

1. Transformation of heterogeneous echocardiographic reports into a structured data format by using Macro VBA.
2. Data cleaning and missing values will proceed before examining exposure outcome associations.
3. Derive normal reference range values for measurements of transthoracic echocardiographic parameters in tertiary hospital of south India.
4. To explore geographic trends in the use of echocardiography in a tertiary care hospital.
5. Develop a prediction model that can predict cardiovascular disease using Statistical Machine learning and Data Mining techniques. [5,8,10]

5. Ethics Approval

Institutional ethics committee of JSS Medical College approved the study and permission was obtained from the hospital to access the dataset from Electronic Health Records (EHRs).

III. RESULTS AND DISCUSSION

Macro VBA algorithm was applied to 16580 records of patients who underwent transthoracic echocardiography at a single academic medical center in the year 2016. The electronic database was in excel sheets containing the ECHO reports for every two months data were provided in each worksheet. A retrospective study was designed to access the echocardiography records from EHRs for a period of one year by using macro visual basic. A total of 55 attributes are commonly evaluated in clinical practice and research studies. These are shown in Table 1. A total of raw quantitative echocardiographic parameters are shown using a Box and Whisker plot distribution according to the outliers. It includes various measurements of the aorta, left ventricular, right ventricular, pulmonary valve chambers of the heart, valves of the heart, wall diameter, and flow velocities across the valves. We present a Macro VBA algorithm capable of large-scale transformation of heterogeneous echocardiographic reports into a structured data format. The feasibility and reliability of an algorithm to transform three categories of data contained within an echocardiography: (i) structured data (ii) semi-structured data and (iii) unstructured data. Generally, structured data are quantitative measures such as wall thicknesses, chamber dimensions, or flow velocities. Semi-structured data fields contain ordinal data. For example, valvular lesions and abnormalities of ventricular function are often subjectively quantified as “mild”, “moderate”, or “severe”. Unstructured fields contain unrestricted prose descriptions of clinically relevant findings as interpreted by the physicians. Researchers and computer scientists have recognized the potential of using unstructured data sources and a major barrier to cover unstructured data to improve patient care is the availability of tools that permit the extraction of high-quality data from unstructured data. Manual extraction of unstructured data from clinical reports is prohibitive, whereas computer algorithms and programming techniques have value for the extraction of information from biomedical science and have an impact on clinical practice and research. The outcomes presented in this paper are the first results from ongoing research from a series of healthcare [13,14,16].

1 Application of the Macro VBA program for the extraction of echocardiographic data.

VBA stands for Visual Basic for Applications, a powerful built-in programming language from Microsoft that is now predominantly used with Microsoft office applications such as MS Excel, MS-Word, and MS-Access. The advantage of using VBA is that we need not install visual basic on our PC, however, installing Office will implicitly help in achieving the purpose. The program design is one of the important things in the technique and information world, therefore, it programming language to design a special program. So, the macro visual basic is used in this research and detect the type of database that must be used to save the desire information needed. Macro VBA is designed to extract semantically viable information to support the heterogeneous clinical research domain and to be sufficiently scalable and robust to meet the rigors of a clinical research production environment. The total time taken by Macro VBA for extracting 16580 echocardiography reports and generating output in a structured format was less than an hour on a personal laptop. The final output consisting of the targeted data elements for extraction with their respective values was generated in a predefined structured format to facilitate further analyses [17].

2. Steps followed to execute the Macro VBA program.

Step1: Open M-S Excel on the PC.
Step2: Access Developer tab option in Excel.
Step3: Open the Visual Basic Editor (VBE) by clicking on “Visual Basic” in the Developer tab or Press. “Alt + F11”. To open VBE.
Step4: Write the VBA code in the “Module1” window.
Step5: Place a command button on the worksheet to run a Macro Visual Basic program.
Step6: The result will be generated in the new worksheet.

3. Data preprocessing

Data preprocessing is imperative in applying data analytics to predict cardiovascular diseases. The study selection process was applied to 16580 of 6891 outpatients and 9689 inpatients transthoracic echocardiography in 2016. 3639 screening ECHO reports, 644 non sequential data, incomplete ECHO reports with missing values, paediatric patient and outliers were excluded and included the normal ECHO records, ischemic heart disease, and other diseases ECHO records are displayed in Fig. 3. During data cleansing, repetitive records, spelling errors and unrealistic data are detected and removed by using Macro VBA. The missing values in the data are identified and replaces them with the mean value of the data. Box and Whisker plot distribution of raw quantitative echocardioatomic parameters are shown in Fig.4. Multivariate identity and extreme outliers are existing in the data series. However, <14 to > 40 (AO), <14 to 38 (LA), <12 to > 25 (RV), <25 to > 58 (L VID_d), <25 to > 40 (L VID_s), >0.6 to > 11 (IVS_d), >13 (IVS_S), >0.6 to > 11 (LVPW_d), <0.6 to > 13 (LVPW_s), >50 (EDV), >30 (ESV), >20 (SV), >60 (EF (%)), >30 (FS (%)), <46 to >112 (MV_E), <35 to >98 (MV_A) these values which lies on either of the extremes can be considered abnormal and should be discarded from the entire series so that any analysis made on this series is not influenced by these extreme values. So, the data series that should be considered for further observation or study after discarding the outliers. Fig. 5. Shows the flowchart of Macro VBA for the extraction of data elements and values into a structured format from structured, semi-structured, and unstructured echocardiography reports [5,18].
Fig. 3. Flow chart for study selection process in echocardiography report at JSS hospital in 2016.

AO, Aortic Root; LA, Left Atrium; RV, Right Ventricle, LVID_D, Left Ventricle Internal Diameter during Diastole; LVID_S, Left Ventricle Internal Diastole during Systole; IVS_D, Intact Ventricular Septum Diameter during Diastole; IVS_S, Intact Ventricular Septum Diastole during Systole; LVPW_D, Left Ventricular Posterior Wall Diameter during Diastole; LVPW_S, Left Ventricular Posterior Wall Diastole during Systole; EDV, End Diastolic Volume; ESV, End Systolic Volume; SV, Stroke Volume; EF (%), Ejection Fraction; FS(%), Fractional Short; MV_E, MV_A, Mitral valve - ratio of the early (E) to late (A) ventricular filling velocities; MR, Mitral regurgitation; TV_E, TV_A, Tricuspid valve- ratio of the early (E) to late (A) ventricular filling velocities; TR, Tricuspid regurgitation; AV_VMAX,
Aortic Valve - The maximal aortic jet velocity; AR, Aortic regurgitation; PV_VMAX, Pulmonary Vascular - The maximal Pulmonary jet velocity; PR Pulmonary regurgitation.

**Fig.4.** Box and Whisker plot distribution of raw quantitative echocardiographic parameters.

![Box and Whisker plot](image)

**Fig.5.** Macro VBA flowchart for extraction of data elements and values into structured format from structured, semi-structured, and unstructured echocardiography reports.

4. **Algorithm**

Step1: Start
Step2: Input the raw data into the worksheet
Step3: Write the VBA code in the “Module1” window by using Visual Basic Editor (VBE) in the Developer tab.
Step4: Add syntax for labels of the attributes and print them into the first row of the result sheets.
Step5: Traverse the sheets row by row using Fn + F8.
Step 6: If (row == Demographical data) split the Date, Month, Time, IP No, Sex, Age and Place and then execute in the separate cell address.

Step 7: Else if (row == M-Mode Measurements) print the measurements value of AO, LA, RV, L VID_d, L VID_s, IVS_d, IVS_S, LVPW_d, LVPW_s, EDV, ESV, SF (%), FS (%)

Step 8: Else if (row == Doppler Measurements) print MV_E, MV_A, MR, TV_E, TV_A, TR, AV_VMAX, AR, PV_VMAX, PR

Measurements values

Step 9: Else if (row == Two-Dimensional Imaging) print Left Ventricle, Left Atrium, Right Ventricle, Right Atrium, Aorta, Pulmonary Artery, IVS, IAS, Mitral valve, Aortic valve, Tricuspid valve, Pulmonary valve, Pericardium, Colour Doppler, Doppler Study, Other

Step 10: Else if (row == Impressions) print the values

Step 11: Repeat the process till end row

Step 12: Stop

After filtering outliers, a manual review was taken by the cardiologist for the cleaning process. The systematic filtering of quantitative echocardiographic data was examined for extreme outliers, unrealistic values, unusual relation and relative values between parameters were identified and data points were retained when found to be valid, edited when obvious data entry errors were identified, and removed in all other cases automatically by using VBA program. The study presents the extraction of heterogeneous or non-standardized transthoracic echocardiographic reports from electronic health records by using the Macro VBA algorithm and solve the problems associated with real time analysis of large-scale data [5,13,14,18].

IV. CONCLUSION

The study presented in this paper is the first ever designed algorithm to linking large scale data from EHRs to transform the information from unstructured into a structured form by using Macro VBA, for investigating the quality of care and improving cardiovascular diseases. The outcome of this study can be used to analyze the geographic trends in the use of ECHO test, derive normal reference value in ECHO parameters for south Indian population and develop a prediction model that can predict cardiovascular disease cases based on measurements taken from transthoracic echocardiography examination using data analytics tools for facilitating data base decision support system.

REFERENCES


