Evaluating the Efficacy of RegEx and R code to Extract Data from Transthoracic Echocardiogram reports

Nischal Praveen, Msc; David Tofovic, MD

Background: The use of electronic health records has allowed for the documentation and preservation of medical information. Beyond its benefits for the patient and healthcare providers, it has been pivotal in the digital era to allow for the collection and analysis of data for clinical research. In cardiology, the Transthoracic Echocardiogram (TTE) is an invaluable imaging modality that provides physicians with the ability to visualize the various anatomical features of the heart and its functioning, informing diagnosis and treatment plans. The TTE, a real time view, is therefore also highly useful in the context of clinical research to better understand various pathological states in patients at varying stages of disease progression, of various ethnicities, ages, and comorbidities. However, there are some challenges in extracting data from TTE reports that results in its underutilization in research. Manual extraction is the gold standard but time consuming and not feasible for large scale investigation where automation is more suitable. Strides have been made to standardize the reporting format of the TTE yet still, natural language discrepancies, inter-author and inter-institutional differences must be considered in developing extraction tools. This study seeks to evaluate the efficacy of regular expressions in R to extract data from reports.

Methods: 1000 echo reports were in CSV form. Of these reports, 26 were removed due to the one of two reasons: 1) Classified as Congenital TTEs 2) Complete absence of data. They were then deidentified using python by removing patient name and MRN. Next, regular expression patterns were diligently developed to capture a wide array of phrases and qualifications for various parameters of interest. Logic checks were incorporated to ensure parameters within the report that were reported multiple times were identical to one another. Data extraction was outputted to a spreadsheet. The following packages were used: readxl, stringr, tidyverse, openxlsx. The code was revisited multiple times to improve the ability to capture parameters even when reported with varying language. 50 reports were manually extracted and compared with the automated data extraction tool. Once the code was sufficiently accurate (>99.7%), the data extraction tool was applied to the larger dataset containing 1000 reports. Thereafter, NA’s (not applicable/empty data points) were analyzed for most variables except those which were rarely expected to be populated in the TTE reports such as MV Stenosis, AV Stenosis, TV Stenosis, PV Stenosis, Simpson’s, TR Velocity, and PASP. Once all NA’s were analyzed, the regular expression patterns were further honed. Subsequent analysis and summary statistics were performed using RStudio.

Results: Out of a potential maximum of 29423 data points, 23080 were obtained. Out of 6343 NA data points, 801 were manually verified. Upon manual verification, 742/801 (92.63%) were correctly NA. 59/801 (7.37%) were found to be incorrectly NA or rather, contained relevant data regarding the specific variable/parameter that should be captured. Date, age, systolic blood pressure, diastolic blood pressure, heart rate, BSA, LVDiD showed 100% estimated accuracy. The variable with the lowest accuracy of NA’s was LA Size at 50.0%. Other variables that were below 90% accuracy of NAs were LV Size (88.2%), LV Hypertrophy (88.9%), LV Function (83.3%), LVEF (87.5%), RV Function (73.1%), MV Regurgitation (88.7%), MV Structure (68.8%), AV Structure (61.8%), and TV Structure (85.7%). All other variables were of higher accuracy (≥99.2%).

Conclusions: Variables and parameters can be reliably extracted from TTE reports using regular expressions in R due to the standardized language and formatting that is used. Numerical parameters follow a consistent pattern in TTE reports and were expected to be extracted with a high degree of confidence. However, this study also demonstrates that text-based appraisal or various variables in a TTE can also be reliably extracted. In future, manual extraction of all 1000 reports will allow for a more thorough evaluation of this tool. Other modalities such as tokenization and natural language processing packages may also be explored.