# Mining textual data in the EMR for prediction of Atrial Fibrillation/Flutter (AFF) through application of machine learning


**Rajesh Chowdhary**[1] CHOWDHARY.RAJESH@MCRF.MFLDCLIN.EDU
**Bess Berg**[2] BESS@CS.WISC.EDU
**Sin Lam Tan**[1] TAN.SINLAM@MCRF.MFLDCLIN.EDU
**Eneida A. Mendonca**[2] EMENDONCA@BIOSTAT.WISC.EDU
**Percy N. Karanjia**[1] KARANJIA.PERCY@MARSHFIELDCLINIC.ORG
**Ingrid Glurich**[1] GLURICH.INGRID@MCRF.MFLDCLIN.EDU
**Romel Garcia-Montilla**[1] GARCIA-MONTILLA.ROMEL@MARSHFIELDCLINIC.ORG
**Peggy L. Peissig**[1] PEISSIG.PEGGY@MCRF.MFLDCLIN.EDU
**Humberto J Vidaillet**[1] VIDAILLET.HUMBERTO@MCRF.MFLDCLIN.EDU
**David C. Page**[2] PAGE@BIOSTAT.WISC.EDU

1Marshfield Clinic-Marshfield Center, 1000 North Oak Avenue, Marshfield, WI 54449, USA

2Department of Biostatistics and Medical Informatics, University of Wisconsin Medical School, 1300 University Avenue, Madison, WI 53706, USA



## Abstract

A large quantity of useful information (e.g. symptoms) is contained in Electronic Medical Record as free-text fields that could potentially be used to build predictive models for disease onset. In this study, we have evaluated the usefulness of free-text based machine learning models in predicting the onset of atrial fibrillation/flutter. Preliminary results of our analysis are encouraging though further testing is required to establish the suitability of our technique in more practical sense.


## 1. Introduction

A new approach employed by researchers for prediction of disease onset is to create predictive models that use coded phenotypic data available in electronic medical record (EMR). Such coded data are typically present in structured forms, (e.g. ICD 9 diagnostic codes), which can be readily retrieved and used for any targeted analysis. However, the strategy of using coded data alone for disease prediction is typically limited. For example, an ongoing Wisconsin Genomics Initiative (WGI) study is seeking to predict atrial fibrillation/flutter (AFF) onset by application of machine learning (ML) models that are trained on coded phenotypic data in the EMR (Berg 2010, Berg et al. 2010). Atrial fibrillation and atrial flutter are the most common cardiac arrhythmias and have been associated with multiple clinical (Greenlee and Vidaillet 2005; Fox et al 2004; Lubitz et al 2009), genetic (Asselbergs et al 2006; Ellinor et al 2008; Lubitz et al 2009; Gudbjartsson et al 2007) and environmental factors (Greenlee and Vidaillet 2005). Presently, using coded data alone, a 60% overall accuracy for disease prediction has been achieved (Berg 2010). In this study Berg developed predictive models trained on coded data including ICD9 diagnoses, laboratory data, vitals, procedures and prescriptions) obtained for AFF patients from Marshfield Clinic's EMR. The models used a combination of rules, each of which is individually informative with a recall on the order of 15% of AFF cases and precision on the order of 70%.

However, a large quantity of useful information (e.g. symptomology) is contained in uncoded and unstructured formats as free-text fields of the EMR that could potentially be used to build improved predictive models of AFF. In this study, we have extended Berg's (2010) AFF predictive modeling on coded-data by evaluating the usefulness of free-text data in modeling AFF onset prediction. We have explored the suitability of using text data for building accurate ML models for predicting the onset of AFF. As a proof of concept, we have developed and tested an ML-based text mining approach for predictive modeling of AFF onset using Marshfield Clinic's textual EMR data. We extracted the textual EMR data associated with the patients we used in our preliminary study (Berg 2010, Berg et al. 2010) and trained ML models with textual features alone in order to


Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

test the accuracy of such models in predicting the onset of AFF. We used several ML techniques for our analysis, such as, Naïve Bayes, Logistic Regression and Support Vector Machines. Of these, Naïve Bayes gave us the best classification accuracy overall, with F-measure of 58.9%; AUROC of 65.9% and overall accuracy of 62.1%. The results of our analysis were encouraging overall suggesting that textual EMR data could be explored for prediction of disease onset.

## 2. Material and Methods

In this study, we used the following case/control definitions as were applied in the previous AFF predictive modeling analysis (Berg 2010), to identify AFF patients in the Marshfield clinic's EMR data:

i) Case definition:

- received at least one diagnosis of either Atrial Fibrillation or Atrial Flutter by a specialist
- AND had an EKG on record
- AND had an annotation (string search) hit for either Atrial Fibrillation or Atrial Flutter (most restrictive definition, at the beginning of the first line of annotations)
- AND did not have surgery (CABG, valve, open or transcatheter procedures for Atrial Fibrillation ablation or internal trauma) within one month of incident diagnosis
- AND have never received a diagnosis of hyperthyroidism.

ii) Control definition:

- lack of arrhythmia diagnoses by cardiologists following review of 12 lead EKG reports.

The control group was matched with our case group by:

- gender
- age (making sure the controls lived to the age at which their matched case had first incidence of AFF)
- birth year (to avoid differences resulting from the history of the EMR, e.g. coded prescriptions were only introduced into the EMR in the 1990's).

Cases were right-censored one week before their first incidence of AFF. Controls were right-censored at the age at which their matched case was right-censored. All records prior to that censor-age were included in our data.

Using the above AFF phenotypic information and matching scheme, all possible text documents available in Marshfield Clinic's EMR after right censoring were extracted for patients identified in our case/control groups. We used all types of textual documents for our analysis. For example, these included, but were not limited to, clinic office notes, interpretations of radiological findings and hospital discharge summary. These documents are largely archived as text files in the 'Documents' section of Marshfield Clinic's EMR. These textual data files are likely to contain information, including AFF symptomology, which is otherwise not captured by coded data in the EMR. Our strategy of focusing on all types of documents rather than only targeted types was to increase coverage of any available information. Moreover, records at the level of individual document types can show high variability in content and presentation over time, which could have confounded our analyses. We term textual data of case/control samples as master text dataset (MTD), and coded data (e.g. ICD9 diagnoses, labs, vitals, procedures and prescriptions) of these case/control samples used in previous analyses as master coded dataset (MCD).

In this study, we analyzed textual EMR data (i.e. MTD dataset) for cases and controls in a temporal window of one year duration prior to (or left of) the reference time point (refer Figure 1 below). In the analyzed range of one year, we identified 546 cases and 545 controls in MTD, as against 679 each for cases and controls in MCD identified in our previous study (Berg 2010). This difference in numbers of cases and controls between MCD and MTD was because certain patients did not have the textual data entry in the target analyzed range of one year.

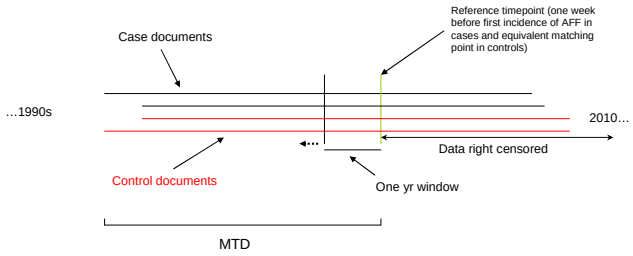


*Figure 1*: Schematic of text data that we analyzed.

### 2.1 Keyword analysis

We performed a text analysis on MTD in our target temporal range to find the informative keywords in MTD that can discriminate cases and controls. For this purpose, we collected a set of keywords that are likely to be associated with AFF disease. This was done by collecting keywords (belonging to signs and symptoms, diseases, drugs, chemical compounds) that were associated with the AFF disease in NLM Medline MESH terms. We then used simple regular expression based matching of keywords in our textual data. We employed word-stemming in our process of keyword recognition. We then analyzed which of our keywords were present or absent in cases and controls of our dataset.

### 2.2 Model Building and Validation

*Training dataset*: We used keywords as features in our predictive modeling. For each keyword, we assigned one of the two values to each of our MTD case/control samples: '0' if the keyword is absent in the sample and '1' otherwise. This dataset of feature vectors for each sample is designated as the 'training dataset'.

*Feature selection*: By applying the method proposed by Hall (1998) on our training dataset with a 10-fold cross validation mode, we found a non-redundant set of keywords that were highly correlated with the class variable. The objective here was to find the most robust subset of non-redundant features from the input training data that might collectively predict AFF-onset with reasonable accuracy. This could help us distinguish the most important features from among a large number of candidate features that might possibly have lot of redundancy between them.

*Model training*: We then used the training dataset along with the short-listed keyword features to train machine learning based classifier models for predicting onset of AFF. The strength of machine learning based prediction models lies in effectively combining a large number of features. Our AFF study involved examination of numerous textual features, and thus ML served as an ideal candidate technique for exploration within this domain. For this study, we explored application of some of the well known ML techniques including: Naive Bayes and logistic regression and Support Vector Machine (SVM). These techniques have previously proved successful in generating reliable predictive models to analyze biomedical data.

We conducted a 10-fold cross validation (a gold-standard in ML (Kohavi 1995)) to evaluate the performance of our trained ML classifier model in discriminating the cases and controls in our target dataset. During the 10-fold cross validation procedure, the data was divided into 10 groups of equal numbers of patients. The learning algorithm then constructed 10 models, each time using only nine-tenths of the data. The model was then tested on the remaining one-tenth of the data. The accuracies were averaged over the ten different runs, each of which employed a different 'tenth of the data' as the test data.

## 3. Results and Discussion

Of all the keywords that we analyzed, 821 keywords found at least one match in our target dataset. Of these, Table 1 shows the list of top 78 keywords that collectively were most correlated with the class variable. Some of the keywords in the list are already known to be associated with AFF. It would be interesting to analyze the association of these keywords with AFF onset in more detail.

Table 2 shows performance of some of the state of the art ML methods on our training dataset based on 10-fold cross validation. We observe that Naïve Bayes outperforms logistic regression and SVM on our dataset in terms of classification accuracy. Naïve Bayes gave an overall accuracy of 62.1%, with F-measure of 58.9% and AUROC of 65.9%.

Overall, the results that we have obtained in our preliminary study are encouraging that merits further investigation in the domain. As a follow up of this study, we intend to explore MetaMap (with Negex) for tagging of keywords in our dataset. In addition, we will also explore keywords that may belong to categories other than those that we used in this study, such as, diet, environmental factors and others. We also intend to analyze temporal windows of other different sizes (e.g. 3 yrs, 5yrs and others).

*Table 1*: Short-listed keywords used for building ML classifier.

| ID | Keyword |
|---|---|
| A36 | Dizziness |
| A42 | exertional dyspnea |
| A303 | Ash |
| A322 | Cardiomyopathy |
| A355 | dilated cardiomyopathy |
| A410 | Liposarcoma |
| A414 | Lymphadenopathy |
| A463 | pulmonary hypertension |
| A469 | renal failure |
| A561 | PFT |
| A605 | Amiodarone |
| A719 | Metoprolol |
| A760 | Refusal |
| A812 | phosphorus |
| A43 | exhaustion |
| A221 | PND |
| A265 | Sustained ventricular tachycardia |
| A292 | aggression |
| A370 | forgetting |
| A535 | Cordarone |
| A657 | Doxorubicin |
| A682 | HCl |
| A685 | Heparin |

| | |
|---|---|
| A713 | Melphalan |
| A724 | Nifedipine |
| A780 | Thrombin |
| A803 | creatinine |
| A178 | INS |
| A227 | Personality Disorders |
| A245 | Reciprocating tachycardia |
| A274 | Tuberculosis Cardiovascular |
| A371 | gammopathy |
| A433 | myopia |
| A483 | spondyloarthropathy |
| A617 | BNP |
| A630 | Chloride |
| A716 | Methacholine |
| A764 | SVC |
| A791 | acetate |
| A26 | chest tightness |
| A177 | ICD |
| A181 | Intermittent Claudication |
| A277 | Vascular Diseases |
| A352 | dermatomyositis |
| A443 | otitis media |
| A444 | papillitis |
| A447 | pericardial effusion |
| A501 | tuberculosis |
| A529 | Cisapride |
| A542 | Etoposide |
| A567 | Prednisone |
| A695 | Insulin |
| A733 | PLA |
| A787 | Vincristine |
| A53 | hypertonia |
| A103 | Arterial Occlusive Diseases |
| A134 | Coronary Stenosis |
| A156 | Factitious Disorders |
| A160 | Glaucoma Open-Angle |
| A179 | Infection |
| A197 | Mediastinal mass |
| A211 | Myxoma |
| A294 | alcohol abuse |
| A323 | cardiovascular disease |
| A326 | cerebral ischemia |
| A327 | cerebrovascular insufficiency |
| A336 | cirrhosis |
| A378 | heart disease |
| A559 | Norepinephrine |
| A564 | Potassium Chloride |
| A600 | Adenosine |
| A663 | Ephedrine |
| A669 | Fibrinogen |
| A203 | Mobitz |
| A332 | chronic ischemic heart disease |
| A513 | ACT |
| A594 | APS |
| A671 | Fluoride |

*Table 2*: Performance of different ML algorithms on our training dataset based on 10-fold cross validation.

| | Precision (%) | Recall (%) | F-measure (%) | AU-ROC (%) | Overall correct predictions (%) |
|---|---|---|---|---|---|
| Naïve Bayes | **64.5** | **54.2** | **58.9** | **65.9** | **62.1** |
| SVM (poly kernel) | 63.9 | 37.5 | 47.3 | 58.1 | 58.1 |
| Logistic | 63.1 | 47.1 | 53.9 | 63.2 | 59.8 |

## References

Asselbergs FW, Moore JH, van den Berg MP, Rimm EB, de Boer RA, Dullaart RP, Navis G, van Gilst WH. A role for CETP TaqIB polymorphism in determining susceptibility to atrial fibrillation: a nested case control study. *BMC Med Genet*., 7:39-47, 2006.

Berg, B. Reducing Overfitting in High-dimensional Relational Rule-Learning. *A preliminary report submitted to the Department of Computer Sciences at the University of Wisconsin-Madison*, Sep 17, 2010.

Berg, B., Peissig, P., Page, D. and Vidaillet, H. Relational Rule-Learning on High-dimensional Medical

Data. *In NIPS Workshop on Predictive Models in Personalized Medicine*, 2010.

Ellinor PT, Yi BA, MacRae CA. Genetics of atrial fibrillation. *Med Clin North Am*., 92:41-51, 2008.

Fox CS, Parise H, D'Agostino RB Sr, Lloyd-Jones DM, Vasan RS, Wang TJ, Levy D, Wolf PA, Benjamin EJ. Parental atrial fibrillation as a risk factor for atrial fibrillation in offspring. *JAMA*, 291:2851-2855, 2004.

Greenlee RT, Vidaillet H. Recent progress in the epidemiology of atrial fibrillation. *Curr Opin Cardiol*., 1:7-14, 2005.

Gudbjartsson DF, Arnar DO, Helgadottir A, Gretarsdottir S, Holm H, Sigurdsson A, Jonasdottir A, Baker A, Thorleifsson G, Kristjansson K, Palsson A, Blondal T, Sulem P, Backman VM, Hardarson GA, Palsdottir E, Helgason A, Sigurjonsdottir R, Sverrisson JT, Kostulas K, Ng MC, Baum L, So WY, Wong KS, Chan JC, Furie KL, Greenberg SM, Sale M, Kelly P, MacRae CA, Smith EE, Rosand J, Hillert J, Ma RC, Ellinor PT, Thorgeirsson G, Gulcher JR, Kong A, Thorsteinsdottir U, Stefansson K. Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nature*, 448:353-357, 2007.

Hall MA. Correlation-based Feature Subset Selection for Machine Learning. *PhD thesis*. U Waikato, Hamilton, New Zealand. 1998.

Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 2(12):1137–1143, 1995.

Lubitz SA, Yi BA, Ellinor PT. Genetics of atrial fibrillation. *Cardiol Clin*., 27:25-33, 2009.

Vidaillet H, Granada JF, Chyou P, et al.: A population-based study of mortality among patients with atrial fibrillation or flutter. *Am J Med.*, 113:365–370, 2002.