34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

# Predicting Atrial Fibrillation and Flutter using Electronic Health Records

Shreyas Karnik[1], Sin Lam Tan[1], Bess Berg[2], Ingrid Glurich[1], Jinfeng Zhang[3], Humberto J Vidaillet[1], C. David Page[2], Rajesh Chowdhary[1*]

*Abstract* —**Electronic Health Records (EHR) contain large amounts of useful information that could potentially be used for building models for predicting onset of diseases. In this study, we have investigated the use of free-text and coded data in Marshfield Clinic's EHR, individually and in combination for building machine learning based models to predict the first ever episode of atrial fibrillation and/or atrial flutter (AFF). We trained and evaluated our AFF models on the EHR data across different time intervals (1, 3, 5 and all years) prior to first documented onset of AFF. We applied several machine learning methods, including naïve bayes, support vector machines (SVM), logistic regression and random forests for building AFF prediction models and evaluated these using 10-fold cross-validation approach. On text-based datasets, the best model achieved an F-measure of 60.1%, when applied exclusively to coded data. The combination of textual and coded data achieved comparable performance. The study results attest to the relative merit of utilizing textual data to complement the use of coded data for disease onset prediction modeling.**

## I. 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 health records (EHR). Such coded data are typically present in structured format, (e.g. ICD 9 diagnostic codes, laboratory tests, vitals), which can be readily retrieved and used for any targeted analysis. However, the utility of using coded data alone for disease onset prediction is typically limited due to missing data values. A large quantity of useful information (e.g. symptomology) is contained in uncoded and unstructured formats as free-text fields within EHR that potentially contains data that could be used to build improved predictive models for disease onset. Making use of the unstructured information in EHR to predict the onset of disease is a challenging problem and an emerging paradigm for opening new perspectives in identifying meaningful secondary use of EHR data.

Atrial fibrillation and atrial flutter (AFF) are the most common cardiac arrhythmias and have been associated with multiple clinical [1-3], genetic [3-8] and environmental factors [1]. In this study, we have explored the suitability of using textual EHR data for building accurate machine learning (ML) generated models for predicting the onset of AFF. The generated predictive models can also be used to phenotype previously uncharacterized AFF patients by utilizing EHR data for a targeted case/control study on AFF, which can complement usual approach of phenotyping patients though expensive manual chart abstraction.

We modeled the problem of predicting the onset of AFF as a classification problem. As a proof of concept, we have developed and tested an ML-based approach for predictive modeling of AFF onset using Marshfield Clinic EHR data. We extracted data associated with AFF cases and matched controls to train and test ML models with features associated with, i) textual-data, ii) coded-data and iii) combination of both textual and coded data in predicting the onset of AFF. We used several classification algorithms such as, naïve bayes, logistic regression, random forests and support vector machines (SVM) along with feature selection to train the ML models. The results of our analysis suggested that textual EHR data should be explored further with coded data for modeling prediction of disease onset.

## II. METHODS

In this study, we used the following case/control definitions, to identify AFF patients in the Marshfield clinic's EHR data:

i. Case definition:
   - received at least one diagnosis of Atrial Fibrillation and/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 AFF diagnoses confirmed by at least one 12 lead EKG interpreted by a cardiologist. 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)

S. Karnik, S.L. Tan, I. Glurich, H.J. Vidaillet and R. Chowdhary (corresponding author, email: chowdhary.rajesh@mcrf.mfldclin.edu, phone: 715-221-6421) are with Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, Marshfield Clinic, 1000 North Oak Avenue, Marshfield, WI 54449, USA
B. Berg and C.D. Page are with Department of Biostatistics and Medical Informatics, University of Wisconsin Medical School, 1300 University Avenue, Madison, WI 53706, USA.
J. Zhang is with Department of Statistics, Florida State University, Tallahassee, FL 32306, USA.

- birth year (to avoid differences in EHR data capture capability over time. For example, while diagnostic coded data have been available electronically since the early 1960's, prescriptions were only introduced into the EHR in the 1990's)

Cases were right-censored one week before their first incidence of AFF. Controls were right-censored at the age atwhich 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 EHR 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. 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). We also extracted coded data from the EHR including: ICD9 diagnoses, laboratory values, vitals, procedures and prescription data for our target samples. We call this data as master coded dataset (MCD) (refer to [9] for more details).

In this study, we analyzed EHR data for cases and controls in a temporal window of one, three, five and all year (all records) durations prior to (or left of) the reference time point (refer Figure 1).

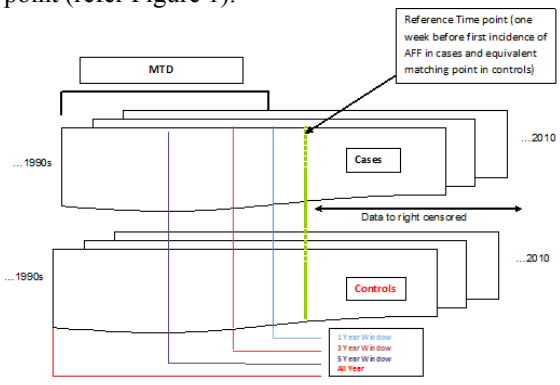


Figure 1: Schematic of EHR data that we analyzed.

We identified 547 cases and 546 controls in MTD and MCD respectively.

*a. Textual EHR data:* When dealing with textual data one challenge is that of extracting features from the text that could be used by the ML algorithms to build classification models. To extract features from MTD we mapped the phrases in the textual data to UMLS [10] vocabulary using MetaMap [11] with the following options: word sense disambiguation, ignore word order and negation detection. We restricted MetaMap mapping to sematic types: drugs, chemicals, diseases and symptoms. By using MetaMap we were able to normalize the variability and ambiguity of terms in the free text. We then generated a non-redundant list of all the UMLS concepts that mapped with the target terms identified in the text. We used identified UMLS concepts as features to train our prediction models. Table 1 shows number of features in each temporal window in MTD that were analyzed.

TABLE I. NUMBER OF FEATURES IN EACH TEMPORAL WINDOW IN MTD SETS

| **Window** | **# of features** |
|---|---|
| Year 1 | 40404 |
| Year 3 | 52252 |
| Year 5 | 57748 |
| All Year | 65850 |

We then used the UMLS concepts present in our MTD samples to generate the following datasets:

*Nominal feature dataset:* For generation of this dataset, we checked for the presence/absence of each UMLS concept in the UMLS mapped sample files and computed a document-term-matrix in which presence of a UMLS concept for a patient sample was encoded as category *1* and absence was encoded as category *0.* We generated individual datasets for four temporal windows and labeled these as MTD1-Year1, MTD1-Year3, MTD1-Year5, and MTD1-AllYear.

*TF-IDF feature dataset:* The approach to generation of the *TF-IDF feature dataset* was analogous to that of the nominal feature dataset but instead of the two-value coding we used 'real value' coding. For each UMLS concept we generated the *TF-IDF* (frequency of UMLS concept/ inverse document frequency of the UMLS concept) features from each MTD temporal windows. TF-IDF is calculated as follows:

$$TF - IDF = tf * idf$$

Where *tf* is the frequency of the term and *idf* is defined as *log* (# of documents/ *tf*). We referred to these datasets as MTD2-Year1, MTD2-Year3, MTD2-Year5, and MTD2-AllYear.

*b. Coded EHR data*: MCD consisted of 53 two-valued features representing coded EHR data including, ICD 9 diagnostic codes, laboratory tests, procedure, medications and vitals.

*c. Combined EHR data*: We combined textual and coded features for each temporal window. We referred to these datasets as MTD1-MCD and MTD2-MCD for each temporal window.

*Feature selection and Model Training:* We have implemented feature selection and ML algorithms in WEKA [12] for this study.

We use a two-step approach to select the most informative features from the textual dataset. In step 1 we first filter features by applying either a Chi-Square filter (threshold:7)

or Information Gain filter (using only top 1000 features) in order to retain only top features that are associated strongly with AFF onset. After applying Chi-Square or Information Gain filters, we applied the CFS feature selection method proposed by Hall [13] to select a subset of features that could strongly discriminate the two classes. We explored different search methods including genetic, linear forward selection, best first, greedy stepwise with CFS to obtain the most relevant features for classification.

We explored application of some of the well-known ML techniques including: naive bayes and logistic regression, random forests and support vector machine (SVM). These techniques have previously proved successful in generating reliable predictive models to analyze biomedical data [14,15].

We conducted a stratified 10-fold cross validation (a gold-standard in ML [16]) to select features and to build predictive models for the onset of AFF. During the 10-fold cross validation process, the data were split into $9/10^{th}$ and $1/10^{th}$ fractions designated as the training fold and testing fold, respectively. We applied feature selection methodology to the training fold and built the ML model using the selected subset of features. We then use the trained ML model to test the performance on testing fold. This process was repeated 10 times. Figure 2 shows a flowchart for this approach. For additional detail with respect to experimental settings and study parameters we used in our analysis, the reader is referred to the following website: http://www.biotextminer.com/affstudy/additionalresults.xlsx

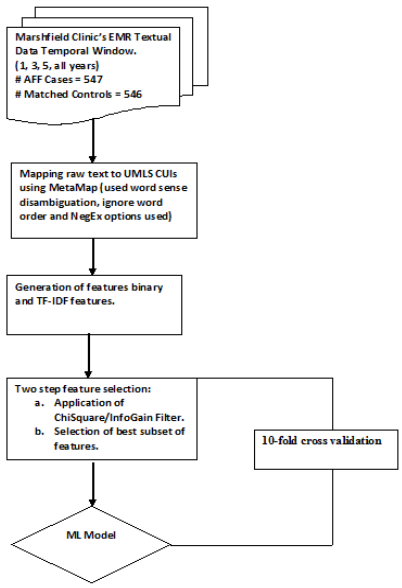


Figure 2: Flowchart of the model building process.

## III. RESULTS

We explored various machine learning algorithms in combination with feature selection methods to obtain accurate AFF prediction models. Table 2 summarizes the performance of these models. Due to space limitations we have only tabulated the performance of the best models in each dataset type MTD1, MTD2, MCD, MTD1-MCD and MTD2-MCD. For detailed results refer to http://www.biotextminer.com/affstudy/additionalresults.xlsx

TABLE II PERFORMANCE OF PREDICTION ALGORITHMS ON DIFFERENT DATASETS USING 10–FOLD CROSS-VALIDATION (ONLY THE BEST PERFORMING TIME WINDOW IS SHOWN FOR EACH DATA SET TYPE). The reader is referred to the following website for a more detailed version of the table: http://www.biotextminer.com/affstudy/additionalresults.xlsx.

| Dataset | NB | SVM | RF | LR |
|---|---|---|---|---|
| MTD1–Year1 | **P:58.6**<br>**R:57.4**<br>**F:57.9** | P:56.1<br>R:51.4<br>F:53.5 | P:57.1<br>R:55<br>F:55.9 | P:58.7<br>R:52.8<br>F:55.5 |
| MTD2–All Year | P:54.2<br>R:51.3<br>F:52.6 | P:58<br>R:58.6<br>F:58 | **P:58**<br>**R:62.7**<br>**F:60.1** | P:56.7<br>R:59.4<br>F:57.6 |
| MCD | P:65<br>R:53.3<br>F:58.3 | P:69.6<br>R:46.6<br>F:55.6 | **P:59.8**<br>**R:61.7**<br>**F:60.6** | P:63.6<br>R:52.8<br>F:57.5 |
| MTD1-MCD –Year1 | **P:59.3**<br>**R:58.8**<br>**F:58.9** | P:56.3<br>R:51.3<br>F:53.5 | P:57.9<br>R:57<br>F:57.4 | P:57.6<br>R:54.1<br>F:55.7 |
| MTD2-MCD–Year 5 | P:56<br>R:50<br>F:53 | P:58.2<br>R:55.5<br>F:56.5 | **P:60**<br>**R:60**<br>**F:60** | P:54.6<br>R:59<br>F:56.5 |

P: Precision, R: Recall, F: F-measure
SVM: Support Vector Machines, NB: Naïve Bayes, RF: Random Forest, LR: Logistic Regression.
MTD1- Textual dataset with nominal encoding
MTD2- Textual dataset with TF-IDF encoding
MCD – Coded Dataset
MTD1-MCD – Combined dataset MTD1 and MCD
MTD2-MCD – Combined dataset MTD2 and MCD

The best performance for AFF onset prediction was achieved in the dataset MTD2-all year window utilizing the random forest classifier. However the performance on MCD dataset and dataset MTD2-MCD (combination of coded and textual data) was also comparable. We also observed that using Chi-Square filter as the feature selection method resulted in better results (Table II) compared to Information Gain filter.

## IV. DISCUSSION AND CONCLUSION

In this study we utilized Marshfield Clinic EHR data to predict the onset of AFF using a combination of feature selection and machine learning approaches. Our results support that textual data can be used and contribute to building reliable AFF onset prediction models.

We observed that on numeric data encoding (MTD2–All Year, MCD) Random Forest classifier out-performed other classifiers, while on two-value categorical data encoding (MTD1–Year1), Naïve Bayes out-performed other classifiers.

On textual data, we found that MTD2 dataset with real-value encoded features performed better than the two-valued encoded feature dataset (MTD1). This was not unexpected since TF-IDF encoding of textual data is expected to be more informative compared to nominal encoding of textual data. Moreover, due to the temporal nature of our data there could be re-occurrence of some symptoms or conditions in the patient records over time. Such aspects could be more effectively captured in the TF-IDF scheme compared to the nominal scheme.

When we compared the performance of coded dataset (MCD) with the textual dataset (specifically dataset MTD2), we found the performance of these two to be similar. This suggests that free EHR text data could potentially be utilized as an inexpensive complimentary method to coded data based approaches for building AFF onset prediction models.

Our study further demonstrated that EHR data can be explored in a temporal manner as demonstrated by our application of temporal windows since disease emergence varies over time. Thus adapting the temporal characteristics of EHR data to optimize building predictive models is a logical approach. We observed that the longer the temporal window the better our outcomes were relative to predicting the onset of AFF. Notably, when the temporal window was longer, richer data were available in the EHR records over time, which could support predictive modeling.

It is necessary to highlight that the annotation/labeling of AFF patient samples in our dataset was based on manual abstraction of patient health records Marshfield Clinic EHR. While we used fairly stringent criteria to identify and label our AFF case samples, in our subsequent error analysis, we identified that in some of the samples the first incidence of AFF was not accurately captured using the criteria we followed for electronic abstraction. For example, first incidence of AFF recorded at Marshfield Clinic may not necessarily represent the first incidence of AFF for the patient, which may be because the patient may have been elsewhere at the time of initial incidence. Such cases possibly led to potential classification errors in our approach.

In conclusion, we selected the most discriminative features for predicting the onset of AFF. We summarized the top 10 features from the overall best performing model (trained on MTD2- all year) in Table III. Further exploration of association of these keywords as AFF onset markers is warranted. For example cigarette (selected best feature) has already been associated with increased risk of AFF [17].

TABLE III. TOP 10 FEATURES SELECTED FROM BEST PERFORMING MODEL.

| CUI | Term |
|---|---|
| C1963220 | Pulmonary hypertension adverse event |
| C2073625 | X-ray of chest: pleural effusion |
| C0340766 | Venous hypertension |
| C0677453 | Magnesium measurement |
| C0373675 | Cigarette |
| C0231819 | Air trapping |
| C2712049 | Actual negative peripheral edema |
| C0699992 | Lasix |
| C0012798 | Diuretics |
| C0202042 | Plasma glucose measurement |

Overall, we demonstrated that prediction of AFF onset using the full-text health records, which was comparable to prediction, achieved using the coded data or the combination of the two data types.

## REFERENCES


[1] R. T. Greenlee and H. Vidaillet, "Recent progress in the epidemiology of atrial fibrillation," *Current opinion in cardiology,* vol. 20, p. 7, 2005.

[2] C. S. Fox, H. Parise, R. B. D'Agostino, D. M. Lloyd- Jones, R. S. Vasan, T. J. Wang, D. Levy, P. A. Wolf, and E. J. Benjamin, "Parental Atrial Fibrillation as a Risk Factor for Atrial Fibrillation in Offspring," *JAMA: The Journal of the American Medical Association,* vol. 291, pp. 2851-2855, June 16, 2004 2004.

[3] S. A. Lubitz, B. A. Yi, and P. T. Ellinor, "Genetics of Atrial Fibrillation," *Cardiology clinics,* vol. 27, pp. 25-33, 2009.

[4] F. Asselbergs, J. Moore, M. van den Berg, E. Rimm, R. de Boer, R. Dullaart, G. Navis, and W. van Gilst, "A role for CETP TaqIB polymorphism in determining susceptibility to atrial fibrillation: a nested case control study," *BMC Medical Genetics,* vol. 7, p. 39, 2006.

[5] P. T. Ellinor, B. A. Yi, and C. A. MacRae, "Genetics of Atrial Fibrillation," The Medical clinics of North America, vol. 92, pp. 41-51, 2008.

[6] D. F. Gudbjartsson, D. O. Arnar, A. Helgadottir, S. Gretarsdottir, H. Holm, A. Sigurdsson, A. Jonasdottir, A. Baker, G. Thorleifsson, and K. Kristjansson, "Variants conferring risk of atrial fibrillation on chromosome 4q25," *Nature,* vol. 448, pp. 353-357, 2007.

[7] J. Granada, W. Uribe, P.-H. Chyou, K. Maassen, R. Vierkant, P. N. Smith, J. Hayes, E. Eaker, and H. Vidaillet, "Incidence and predictors of atrial flutter in the general population," *J Am Coll Cardiol,* vol. 36, pp. 2242-2246, December 1, 2000 2000.

[8] H. Vidaillet, J. F. Granada, P. o.-H. Chyou, K. Maassen, M. Ortiz, J. N. Pulido, P. Sharma, P. N. Smith, and J. Hayes, "A population-based study of mortality among patients with atrial fibrillation or flutter," *The American Journal of Medicine,* vol. 113, pp. 365-370, 2002.

[9] B. Berg, D. Page, P. Peissig, and H. Vidaillet "Relational Rule-Learning on High-dimensional Medical Data" in NIPS Workshop on Predictive Models in Personalized Medicine, 2010

[10] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology," *Nucleic Acids Research,* vol. 32, pp. D267-D270, January 1, 2004 2004.

[11] A. R. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program," in Proceedings of the AMIA Symposium, 2001, pp. 17-21.

[12] E. Frank, M. Hall, G. Holmes, R. Kirkby, B. Pfahringer, I. H. Witten, and L. Trigg, "Weka Data Mining and Knowledge Discovery Handbook," O. Maimon and L. Rokach, Eds., ed: Springer US, 2005, pp. 1305-1314.

[13] M. A. Hall, "Correlation-based feature selection for machine learning," The University of Waikato, 1999.

[14] Q. Zeng, S. Goryachev, S. Weiss, M. Sordo, S. Murphy, and R. Lazarus, "Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system," *BMC Medical Informatics and Decision Making*, vol. 6, p. 30, 2006.

[15] E. Çomak, A. Arslan, and İ. Türkoğlu, "A decision support system based on support vector machines for diagnosis of the heart valve diseases," *Computers in Biology and Medicine*, vol. 37, pp. 21-27, 2007.

[16] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," in *IJCAI*, 1995, pp. 1137-1145.

[17] Heeringa, J. A. Kors, A. Hofman, F. J. A. van Rooij, and J. C. M. Witteman, "Cigarette smoking and risk of atrial fibrillation: The Rotterdam Study," *American Heart Journal,* vol. 156, pp. 1163-1169, 2008.