Personalized heart failure severity estimates using passive smartphone data

Ayse S. Cakmak, Georgia Institute of Technology
Erik Reinertsen, Georgia Institute of Technology
Herman A. Taylor, Georgia Institute of Technology
Amit J. Shah, Emory University
Gari D. Clifford, Emory University

Conference Name: 2018 IEEE International Conference on Big Data
Publisher: IEEE
Conference Place: Seattle, WA, USA
Publication Date: 2018-12-10
Type of Work: Conference | Final Publisher PDF
Permanent URL: https://pid.emory.edu/ark:/25593/tj25t

Final published version: http://cci.drexel.edu/bigdata/bigdata2018/

Accessed May 15, 2021 3:53 AM EDT
Abstract—Heart failure (HF) is one of the leading causes of mortality in the United States with a high economic burden due to readmissions. We present a novel approach to remotely monitor quality of life in patients with HF using a smartphone app and a scalable cloud-based architecture. In a preliminary study, we assess continuous data from 10 HF subjects over a period of up to a year. Over 680 million samples of physical motion data, 9,000 geographic location updates, and 11,000 individual contact activity/diversity events in the form of phone calls were captured from the app. Personalized models were constructed from these data to estimate self-reported quality of life using the Kansas City Cardiomyopathy Questionnaire (KCCQ), which has been shown to be a reliable health status measure for HF patients. Generalized linear models using only motion features were shown to reliably estimate the KCCQ score with an out of sample mean absolute error of 5.71%. Personalized models for estimating the HF severity as mild or severe were also built as a proof of concept to detect when a subject’s data indicated a clinical deterioration. Average out of sample accuracy was 83% for this binary classification problem. Creation of personalized models from passive smartphone data collected ‘in-the-wild’ to identify changes in HF severity appears possible. This new approach holds promise as a low burden and accurate method of monitoring HF symptoms, which could aid clinicians in early assessment and prevention of adverse outcomes.

Keywords—Heart failure; mobile health (mHealth); regression; actigraphy; geolocation; social interaction

I. INTRODUCTION

Heart failure (HF) is characterized by structural or functional cardiac abnormalities that lead to elevated intracardiac pressures and/or reduced cardiac output [1]. HF is one of the most common causes of hospitalization and readmission. In 2011, the HF hospitalization rate in the USA was 18 per 1000 for those over age 64 [2], making HF among the leading causes of hospitalization in this age group [3]. HF readmissions are perceived as a correctable source of poor quality of care and excessive medical spending. The cost of medical care and productivity losses in 2012 due to HF was estimated at $31B, a figure estimated to rise to $53B by 2030 [4]. Since passage of the Affordable Care Act in 2010 in the United States, providers have been financially penalized for higher than average HF readmission rates, and readmissions have declined from 17.2 to 16.1 per 100 hospital discharges [5]. However, one-fifth of those who are hospitalized with HF are still readmitted within 30 days.

Interventions to reduce HF readmissions have only achieved limited success. Telemedicine programs have not been shown to definitively improve morbidity and mortality in HF, with conflicting reports in the literature [6]–[8]. Factors associated with readmission that were assessed in these studies included blood pressure, body weight, heart rate and symptoms, none of which have been shown to be good predictors of acute decompensation [9], [10] even when coupled with an intervention [8].

Smartphones and wearable devices can detect alterations in locomotor activity and other behavioral patterns that may be associated with worsening symptoms such as lower extremity edema, shortness of breath, and fatigue. Previous studies have demonstrated that passively collected activity and location data can be used to approximate self-reported clinical scales, or classify illness severity [11]–[14], and could also be applied to the problem of HF severity estimation. Tang et al. [15] evaluated the relationship between accelerometer-measured activity, survey data, and performance-based function tests in 52 subjects with severe aortic stenosis, a severe cardiovascular disease that limits quality of life and function. Using multiple linear regression models the authors did not demonstrate a relationship between locomotor activity and survey data or function tests. This may have been due to the population-based approach of the work, and perhaps the relatively simple models employed. Estimation of outcomes such as quality of life in HF from passively monitored data gathered via smartphone has not been previously reported. Such an approach could improve monitoring efforts and further our understanding of the physiological and behavioral determinants of HF-related quality of life and predictors of clinical decompensation.

In this work we demonstrate the feasibility of predicting quality of life of patients with HF using passive smartphone data measured via the latest version of our bespoke app [16].
We construct patient-specific models to estimate Kansas City Cardiomyopathy Questionnaire (KCCQ) scores (a validated survey to assess quality of life, predict readmissions and mortality in HF patients) using various features from past passively monitored data, and use these features to cluster patients into high ($KCCQ \leq 25$) versus moderate ($KCCQ > 25$) severity.

II. METHODS

A. Study Architecture

The bespoke “Automated Monitoring of Symptom Severity” (AMoSS) app, designed and implemented by our research team and collaborators, passively collects location, activity, clinical surveys and contact activity/diversity data via de-identified lists of word type, as well as recipients and senders of text messages and phone calls, including length/duration and time of day [16]. In order to protect subject’s privacy, all data are de-identified at source, using hashed identifiers and random geographic offsets. The app uploads data every few hours to Amazon Web Services for storage.

B. Data Collection

Subjects with HF were enrolled in an ongoing HF study at the Veterans Affairs Medical Center in Atlanta, USA. To be included in the study we required at least 80 days of data for each patient. A total of 10 (male only) patients satisfied this criterion to-date. The population age ranged from 59-76 years, and weight from 79-109 kg at the time of collection. The AMoSS app was installed on subject’s phone after they visited the HF clinic. The subject could also elect to stop the collection of individual data types at any time.

Passive data types utilized in this study were: activity, location, and contact activity. Activity data were collected by the smartphone’s 3D accelerometer. Accelerometry was measured in the x, y, z axes and sampled at 5 Hz. Fig.1(a) is a “double plot” that shows activity levels measured via accelerometry over a 150-day period. Each column is created by stacking two consecutive days of data. The first column shows activity levels on day 1-2, the second column shows day 2-3, etc. White regions in the plot indicate a lack of activity data, i.e. if the subject chose to not share data during that time, or if the smartphone ran out of battery. Location data was retrieved using the Android location services application program interface. Fig. 2 shows an example of a subject’s location data. To create a location data entry, the subject was required to move at least 100 meters, and 5 minutes must have passed since the last location data update. Anonymized call data and call duration were used to monitor patients’ contact activity over time (Fig.1).

KCCQ scores were collected from users as an “active data type”, i.e. requiring the user to input values into various fields. The KCCQ score quantifies physical function, symptoms, social function, self-efficacy and knowledge, and quality of life; it is a valid and reliable health status measure for HF patients. A higher score indicates better health status [17]. In this study, we used the KCCQ-12 summary score, which is a shorter but generally equivalent version of the KCCQ [18]. The summary score ranges from 0-100 and is calculated by averaging physical limitation, symptom frequency, quality of life and social limitation domains of questionnaire.
Figure 2: Example of location data collected with AMoSS app. Increasing height, represents the probability of visiting an area. Areas south and west from the origin are represented as negative distances. The red circle is the boundary of “zone-1”, the area enclosed by a circle of 20 km radius from most frequently visited location.

C. Preprocessing

Borazio et al. describes the conversion of raw 3D accelerometer to activity counts [19]. Activity counts are the output format of most commercial actigraphy devices; data is summarized over 30 second epochs or time intervals. This conversion compresses information, reduces required memory for storing data, and eliminates artifacts and noise in raw data. Z-axis actigraphy data was filtered using a 0.25 – 11 Hz passband to eliminate extremely slow or fast movements [20]. The maximum values inside 1-second windows were summed for each 30-second epoch of data. These summations were scaled to obtain activity counts for each epoch [21].

Dates of all KCCQ data from each subject were determined. Data in the two-week window leading up to each KCCQ score were isolated. Because the KCCQ-12 assesses quality of life and health status in the two-week period prior to the questionnaire, a window size of two weeks was chosen. Features were extracted from passively gathered data in this two-week window. Sufficient data within a two-week window was defined as a) at least seven days of activity data and b) at least one upload of location and contact activity data.

D. Feature Extraction

1) Motion: Several features were extracted from movement data to evaluate rest-activity characteristics and circadian rhythms of subjects.

- Rest activity metrics: Rest-activity rhythms were assessed using Interdaily Stability (IS), Intradaily Variability (IV), Most Active 10 Hours (M10) and Least Active 5 Hours (L5) [22]. IS quantifies invariability between days and can be calculated as
  \[ IS = \frac{n}{p} \sum_{h=1}^{24} (\bar{A}_h - \bar{A})^2, \]  
  where \( \bar{A} \) is the mean of all data, \( \bar{A}_h \) are the hourly means, \( A_i \) are the data points, and \( n \) is the number of data points in the interval. IV quantifies fragmentation of the rest-activity rhythm as
  \[ IV = \frac{n}{(n-1) \sum_{h=1}^{24} (A_i - A_{i-1})^2}. \]  

M10 is the average of the most active 10 hours over all days. A drop in M10 could imply a reduction in physical activity due to HF symptoms. Lastly, L5 is the average of the least active 5 hours. L5 indicates movements during sleep, and night-time arousal.

- Cosinor rhythmometry: To extract cosinor rhythmometry features, a cosine model of the following form was fit to the data; \( Y(t) = M + K \cos(2\pi t/\tau + \phi) + e(t) \), where \( M \) is mesor (baseline activity of subject), \( K \) is amplitude (how active subject is during the day versus night), and \( \phi \) is acrophase (a metric of the circadian cycle) [23].

- Auto-correlation coefficient: The correlation coefficient \( R_k \) quantifies the correlation of a variable with itself at a previous time, e.g. \( A_i \) versus \( A_{i-k} \). In this work a lag \( k = 24 \) hours was used. A more pronounced circadian rhythm will result in a higher \( R_k \) [24].

- Total energy of activity data

- Peak of activity: The time of day of the maximum activity rhythm was determined and expressed in radians.

- Multiscale entropy (MSE): MSE was calculated to quantify irregularity or unpredictability of behavior over multiple timescales. MSE was calculated following the methods described by Costa et al. [25]. Actigraphy time series were coarse-grained by averaging the data points within non-overlapping windows. The first 20 scales of multiscale entropy were calculated by varying the window size from 1 to 20. For each coarse-grained time series, sample entropy was calculated.

2) Location: Severe HF causes discomfort and can hinder physical activity, which could lead to the subject staying at home more, or altering routine behaviors. Location features were extracted to capture these changes.

- Percentage of time at “home” location: Using all location data from each subject, the “home” location was defined as the most frequently visited location. The percentage of time spent at “home” was calculated.

- Percentage of time at second most frequent location: The second most frequently visited location was determined in a similar fashion as “home”. The percentage
3) Contact Activity: HF symptoms could affect a subject’s social behavior. The following contact activity features were extracted from smartphone data:

- **Total number of calls**
- **Mean duration of calls**
- **Standard deviation (std) of duration of calls**
- **Mean duration without any calls**
- **Standard deviation of duration without any calls**

### E. Generalized Linear Model for KCCQ Estimation

Personalized models were created for each subject to estimate the KCCQ summary score. Generalized linear models (GLM) with binomial distribution and logit link was used. Elastic net regularization was applied to personalized GLMs to decrease bias and improve classification performance [26]. Models were built only for subjects from whom sufficient data was gathered over at least ten windows, where a window was defined as two weeks. If insufficient data was gathered from ten or more windows, or did not share a specific data type at all, no model was built for that data type.

Model performance was assessed via record-wise leave-one-out cross validation (LOOCV). Given N windows of data, N − 1 windows are used to train the model for a given patient for predicting KCCQ scores, and the held-out window is used as the test set data from which a KCCQ score is predicted. This is repeated for the remaining N − 1 windows. The Mean Absolute Error (MAE) between actual and estimated KCCQ scores was calculated for each personalized model.

When the number of intervals with combined data types was less than five, regularization was applied to most frequently selected features in single data type analysis.

### F. K-Nearest Neighbors Classification

In addition to regression analysis, a classification analysis with a K-Nearest Neighbors (K-NN) approach was performed after quantizing KCCQ summary scores. Data were dichotomized into KCCQ scores ≤ 25 or > 25 [17]. KCCQ scores ≤ 25 correspond to New York Heart Association (NYHA) class IV. Patients with class IV HF are unable to complete any physical activity without discomfort. KCCQ scores > 25 correspond to NYHA class I-III which describes less severe HF compared to class IV. Clustering analysis was performed for two subjects who had enough KCCQ summary scores in each class. Cosine distance and five nearest neighbors were used as model specifications. Five-fold cross validation (CV) was implemented whereby the model was trained on four folds and the fifth held-out fold was used for testing, and this process was repeated for the remaining four folds. The percentage of correctly classified points were reported for each subject.

### III. Results

1) **KCCQ Estimation**: The average MAE over the population for estimating the KCCQ using only activity metrics was 5.71 units (or percent, since the scale is normalized to be between 0-100) as shown in Table I. When only location metrics were used to estimate the KCCQ scores, the MAE rose to 7.40 (N=8). For personalized models based on contact activity features, the population average (N=9) MAE was 6.05. For personalized models built with features from all three data domains, the population average MAE was 5.43. Although one might therefore be tempted to infer that movement provides the most information in this context, inspection of Table I shows that the most useful type of data depends on the individual. Moreover, the most frequently selected features by elastic net in the LOOCV procedure also varied according to the subject. This creates a strong case for personal models trained on initial KCCQ reports. It can also be noted that the models developed here outperform the baseline (sample-and-hold) estimate, i.e. simply using the first KCCQ score. No error exceeded 8% when all three domains of data types were available.

### Table I: Mean absolute error of LOOCV KCCQ estimation models for each of the 10 patients.

<table>
<thead>
<tr>
<th>Initial KCCQ Score</th>
<th>MAE of Motion Features</th>
<th>MAE of Location Features</th>
<th>MAE of Contact Act. Features</th>
<th>MAE of Combined Features</th>
<th>MAE of Baseline Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>95</td>
<td>0.43</td>
<td>0.90</td>
<td>0.91</td>
<td>0.44</td>
<td>2.12</td>
</tr>
<tr>
<td>50</td>
<td>†</td>
<td>2.26</td>
<td>2.47</td>
<td>2.25</td>
<td>40.39</td>
</tr>
<tr>
<td>13</td>
<td>4.93</td>
<td>†</td>
<td>4.01</td>
<td>3.92</td>
<td>4.74</td>
</tr>
<tr>
<td>56</td>
<td>5.45</td>
<td>7.73</td>
<td>6.44</td>
<td>5.45</td>
<td>34.86</td>
</tr>
<tr>
<td>71</td>
<td>5.46</td>
<td>5.44</td>
<td>5.68</td>
<td>5.49</td>
<td>6.61</td>
</tr>
<tr>
<td>38</td>
<td>7.50</td>
<td>12.08</td>
<td>7.41</td>
<td>7.52</td>
<td>17.96</td>
</tr>
<tr>
<td>77</td>
<td>9.41</td>
<td>9.55</td>
<td>7.82</td>
<td>7.85</td>
<td>19.87</td>
</tr>
<tr>
<td>19</td>
<td>†</td>
<td>12.32</td>
<td>8.82</td>
<td>7.99</td>
<td>6.10</td>
</tr>
<tr>
<td>13</td>
<td>8.22</td>
<td>8.31</td>
<td>†</td>
<td>8.33</td>
<td>19.31</td>
</tr>
<tr>
<td>33</td>
<td>†</td>
<td>†</td>
<td>10.90</td>
<td>†</td>
<td>13.32</td>
</tr>
<tr>
<td><strong>AVG</strong> → 5.71</td>
<td>7.40</td>
<td>6.05</td>
<td>5.43</td>
<td><strong>16.53</strong></td>
<td></td>
</tr>
</tbody>
</table>

† indicates patient did not share data type

2) **KCCQ Classification using K-NN**: Using contact activity features from two subjects (Subject 1 and Subject 8 from Table I), five-fold cross validation of the K-NN...
classification approach was performed as described in II-F. Out-of-sample classification accuracy for these two subjects were 0.78 and 0.88 respectively. Repeating the same analysis with location features resulted in respective classification accuracies of 0.65 and 0.73.

IV. Discussion

In this paper, we present a novel approach for monitoring HF patients with only passively data collected from via smartphone. Data collected from smartphones can provide a high-compliance signal which provides valuable insight into a subject’s daily behaviors and health status. We hypothesized that activity, location and contact activity data were all affected by changes in health status for patients with HF, and these data gathered via smartphone could be used to assess patient quality of life. Activity data can be used to infer circadian rhythm of subject, activity-rest schedules and disruptions (such as awakenings during night). For instance, it can be observed from Fig.1(a) the subject is less active than normal before reporting his lowest KCCQ scores. Location data can indicate if subject is disinclined or unable to leave the house or continue a normal routine. Contact activity data provides information about subject’s social interaction, particularly with a given subgroup of contacts. During health changes we observe a change in social behavior with a user changing the frequency or duration or calls, and to whom they are placed, as seen in Fig.1(b). Although none of these changes are particularly specific to changes in HF when taken in isolation, when combined together they provide a more accurate measure of the changes in health.

Features extracted from the three domains (activity, location and contact activity) were used to build clinically useful models. Personalized models were implemented instead of population models because KCCQ scores are subjective and self-reported, and our sample size was too small to generalize a group model to a larger cohort. A total of 30 features were derived from activity data, and five features were extracted from both location and contact activity data. By using only activity data, the MAE in the KCCQ estimate was 5.71%. A KNN model which classifies KCCQ scores as ≤ 25 or > 25 was implemented to detect clinically significant changes in the population. This binary classifier exhibited an accuracy of 78% and 88% for the two subjects who had a sufficient number of passive data and KCCQ scores. Since this study is ongoing, this classification approach will be revisited when more subjects with varying KCCQ levels are enrolled for sufficient time in the study.

We also note that subjects were given the option to determine which types of data to share, and the frequency of their uploads. This option was provided to empower users to take ownership of their data and decide when and how they were monitored, and hence increase compliance. This led to some missing data in some patients. For patients with high mortality risk, these options could be removed to improve data continuity. It is important to note that adherence to uploading of active data sources by patients can decline rapidly over time [27].

We also observed variation in compliance measured by active data uploading, but this was not a consistently diminishing level as implied in other studies. For example, a subject enrolled for 428 days sent passive activity data for 328 days, although only 46 KCCQ survey reports were completed in the same time period. However, we may only need the first few weeks of reports – during the higher compliance period – to build individualized models. In the cases where we have insufficient data to build a model, or the confidence in the model drops below a given threshold, incentives can be provided to report the KCCQ. In this way, we may maximize information and minimize patient burden.

V. Conclusion

This paper describes the use of a bespoke app and a scalable cloud-based system for monitoring HF patients non-intrusively during their daily lives. Models were built to estimate the KCCQ scores – a measure of quality of life in patients with HF – by extracting features from only passive collected data sources. The out of sample accuracy of 83% may be high enough to identify significant changes in severity which would require an intervention or reassessment of treatment. This framework could be applied to a large population with very low cost, high compliance, and a rapid response rate compared to existing approaches.

VI. Acknowledgments

The authors wish to acknowledge the support of the National Science Foundation Award 1636933, NIH/NHLBI award K23 127251 and the Georgia Research Alliance.

VII. References


