TRIPOD Protocol: Development and internal validation of a NYHA functional class prediction algorithm for patients with (severe to moderate) heart failure

Jonathan-F. Baril

October 12, 2017

Abstract

OBJECTIVE: To develop and validate an NYHA functional class prediction algorithm for patients with (severe to moderate) heart failure
DESIGN: Sets, composed of one model for each target NYHA class (II & II), of multivariate hidden (pure- or semi-)Markov models
SETTING: Tertiary care clinic specializing in the management of heart failure (The Heart Function Clinic at Toronto General Hospital, a part of the University Health Network (UHN) in Toronto, Canada) from November 2017 to April 2017.
PARTICIPANTS: consecutive sample of n consenting heart failures patients participating in program evaluation and quality improvement efforts for Medly, a telemonitoring platform used and developed for the test site. [Development; Internal Validation]
PRIMARY OUTCOME: NYHA functional classification as assessed by patient's attending cardiologist at clinic visits during study period.
RESULTS: TBD
CONCLUSIONS: TBD

[Prognosis; Development; Internal Validation]

1 Introduction

1.1 Background & Objectives

Heart failure, the complex chronic terminal phase of all cardiovascular disease, is slowly becoming a worldwide silent pandemic [https://www.ncbi.nlm.nih.gov/pubmed/26462110]. The symptoms of heart failure are complex and difficult to manage for both patients and their physicians. Care is made even more difficult because there is no reliable objective method for assessing the moment-to-moment state of given patients HF, never mind determining if it is likely to worsen though certainly not for lack of possible metrics. The current clinical gold standard for communicating a patients state of Heart Failure is the New York Heart Association (NYHA) functional classification. This system grades a patient's degree of heart failure based on the physicians interpretation of patient reported symptoms (mainly with respect to their degree of exercise/activity intolerance). Despite these limitations clinical evidence and medical research have established many important relationships between a patient's NYHA class and their prognostic outcomes.

Finding an objective means of determining a patient's NYHA class would provide a great boon to both heart failure care and research as it would cause intra- and inter-physician and patient assessments of heart failure class to more consistent. Doing so would make communication of patient heart failure states in research, clinic notes, or other medical documentation more transparent and reliable.

Subjectivity in the current NYHA classification is introduced through two primary sources: patients and clinicians. A continuously worn activity tracker and an intelligent classification algorithm executed by a computer could feasibly replace these two sources and provide a consistent and purely objective method of assessing NYHA class.

Modern commercially available activity/fitness trackers, can fairly reliably track minute-by-minute step count and heart rate, promising to provide a human memory-independent and precise (even if not a 100% accurate) picture of a patient's response to activity. “Replacing” patient memory with activity trackers would eliminate a significant source of subjectivity and potential error in trying to determine a patient's level of activity intolerance.

An intelligent classification algorithm could then be trained to mimic expert grading by an experienced “model” physician (and to compensate for inherent biases and inaccuracies in the recording sensors). By imbuing artificial intelligence into such an algorithm so that it could translate relevant data into the desired clinical outcome (NYHA classification) or a sufficiently equivalent outcome (an NYH-AI or NYHAI classification) we could provide a way for to assess a patient's functional classification in an objective, universally-consistent manner that still leverages the advantages and benefits of the existing 'traditional' NYHA classification method.

2 Methods

2.1 Source of Data/Study Design

The data used for algorithm development was sourced between November 2017 and April 2017 from an open (prospective) cohort of adult outpatients at a tertiary care clinic specializing in the management of advanced heart failure (The Heart Function Clinic at Toronto General Hospital, a part of the University Health Network (UHN), in Toronto, Canada).

The same data used for algorithm development was used for internal validation of the algorithm. No additional sufficiently large dataset was available for external validation at the time of the study.

2.2 Participants

The cohort consisted of all consecutive patients prescribed, enrolled and using (i.e. being monitored by) the Medly telemonitoring program (Medly) consenting to be included in program evaluation and quality improvement efforts for Medly.

A patient was considered for inclusion into Medly only if they were:

1. a consenting adult (18+ years of age),
2. diagnosed with heart failure,
3. followed by a licensed cardiologist at the UHN Heart Function Clinic who in turn bore the primary responsibility for the management and care of that patients heart failure diagnosis (the responsible or attending cardiologist),
4. sufficiently capable of speaking and reading English, or having an informal caregiver (spouse, parent, etc.) capable of the same so as to both:
   (a) undergo the process of and provision of informed consent for participation in the Medly program
   (b) understand and follow the text prompts provided by the Medly patient-side application
5. capable of complying with the use of Medly (e.g. capable of truthfully answering symptom questions, capable of safely and correctly using the peripherals such as the weight scale, activity tracker and blood pressure cuff)

While there were no explicit exclusion criteria for the study, we note that the decision to prescribe or exclude a patient from the Medly program was ultimately up to the professional judgment of the attending cardiologist. During the study period a total of 6 attending cardiologists prescribed Medly as part of patient care although one of the cardiologists (the medical director of the clinic) was disproportionately responsible for a majority of the patients monitored (X/N or Y%).

During the study period patients received no special treatment with regards to the medical interventions used and prescribed to them. Their cardiologist continued to treat them according to the (evolving) standard-of-care at the Heart Function Clinic, which also included regular updates and bug-fixes to the Medly platform.

2.3 Outcome

The outcomes of interest are the New York Heart Association (NYHA) functional class for each patient. While competing methods of reporting the severity of heart failure experienced by a patient exist, the NYHA classification is arguably the most commonly used system for functional classification of heart failure and many well established relationships have been identified between a patient's NYHA class and prognostic outcomes for the condition. NYHA class is assessed for each patient as part of each visit at the Heart Function Clinic. The assessment is performed by the physician seeing the patient (resident, fellow or otherwise) and verified by the attending physician. Within the context of this study it means that outcomes measures are recorded at the date of on-boarding and then only recorded sporadically during the study period as patients return for clinic visits. The interval between regularly scheduled clinic visits varies between 1-2 weeks to 3-6 months. Patients with more severe or less stable conditions will visit more frequently. As well, patient's who are undergoing a change in their heart failure medication (e.g. up- or down- titrating) will also visit more frequency, usually a few times within a period of a few weeks as the medication is adjusted. Since patients were on-boarded on a rolling basis over a period of 5 months there is no guarantee that a given patient (especially the more stable patients) will return to physically visit the clinic within the study period and be reassessed for any potential change in their NYHA class.

If NYHA functional class is too unstable, the Medly symptoms questionnaire could be used as a proxy outcome to detect suspected drops/aggravations in NYHA class. Every morning patients are prompted by the telemonitoring system to answer the following series of questions to ascertain their current condition. Note that questions 7-11 only appear if a patient has answered yes to any of question previous questions (1-6).

1. Have you fainted?

2. Has your ICD (Implantable Cardiverter Defibrillator) gone off?

3. Has your breathing as night worsened?

4. Do you have more chest pain than usual?

5. Are you more tired than usual?

6. Are you more short of breath than usual?

7. Are your ankles swollen?
8. Do you feel that your heart is beating unusually?
9. Do you feel lightheaded?
10. Did you have to stop any of your usual daily activities because of your health?
11. Enter the number of pillows you used to sleep.

While all of the questions target some specific symptom or set of symptoms experienced by heart failure patients, questions 1, 2, 4-6, & 8-10 specifically try to determine whether there have been an increase in a patients exercise/activity intolerance. Since a patient's level of activity intolerance is the primary determinant of their NYHA classification, an increase in the frequency of affirmative responses to these questions over a sustained period, say 1-3 weeks, would be a strong indicator that a patient's NYHA class may have changed. Baseline response frequency can be determined from the 2 week period immediately following the first clinic visit moving to the 2 week period immediately preceding the last clinic visit for all following visits.

2.4 Predictors

The predictors used to generate the prediction outcome are as follows:

1. Heart Rate [beats per minute] - recorded using commercially available activity-tracker (Fitbit Charge HR, Charge 2 or Alta HR) continuously throughout the day
2. Step Count [steps per minute] - recorded using commercially available activity-tracker (Fitbit Charge HR, Charge 2 or Alta HR) continuously throughout the day
3. Blood Pressure [systolic/diastolic in mmHg] - recorded using a bluetooth enabled blood pressure cuff (A&D Medical UA-651BLE or Omron 10 Series BP786N) once daily in the morning, and then at will by the patient or if symptoms worsen during the rest of the day
4. Weight [kg] - recorded using a bluetooth enabled smart scale (A&D Medical UC-352BLE or Ivation IVA-BTS351-B) once daily in the morning, and then at will by the patient or if symptoms worsen during the rest of the day
5. Symptom Questionnaire Responses (when not used as surrogate outcome) - recorded using a smartphone (iPhone Y or Samsung Galaxy Grand Prim) running Medly app version Y.X+ or Z.X+ (respectively) once daily in the morning, and then at will by the patient or if symptoms worsen during the rest of the day. Specifically of interest are the responses to the following subset of questions:

(a) Have you fainted? [yes or no]

(b) Has your Implantable Cardioverter-Defibrillator (ICD) gone off? [yes or no]

(c) Do you have more chest pain than usual? [yes or no]

(d) Are you more tired than usual? [yes or no]

(e) Are you more short of breath than usual? [yes or no]

(f) Do you feel that your heart is beating unusually? [yes or no]

(g) Do you feel lightheaded? [yes or no]

(h) Did you have to stop any of your usual daily activities because of your health? [yes or no]

6. Patient Demographic/Meta Data - recorded when prescribed Medly (carried over from onboarding process before this study), namely:

   (a) Sex [Male or Female]

   (b) Age [years]

   (c) Height [centimeters]

   (d) Handedness [left or right]

   (e) Wristband Handedness Preference: [left or right]

   (f) (Left Ventricle) Ejection Fraction: [%]

   (g) Heart Failure Diagnosis Date [time since in months]

7. Heart Failure Treatment to Date - recorded when prescribed Medly (carried over from onboarding process before study), namely:

   (a) Lifestyle changes [yes or no]

   (b) Implantable Cardioverter-Defibrillator (ICD) [yes or no]

   (c) Left Ventricular Assist Device (LVAD) [yes or no]

   (d) Heart transplant [yes or no]

   (e) Other surgical intervention [yes or no]

8. Prescribed Medications - recorded at onboarding and updated as required at every clinic visit, sourced from clinic notes:

   (a) I_f Channel Blocker [mean daily dose in mg]

   (b) Beta Blockers [mean daily dose in mg]

   (c) Aldosterone Antagonists [mean daily dose in mg]

   (d) Diuretics [mean daily dose in mg]

   (e) Digoxin [mean daily dose in mg]

2.5 Sample Size

As a general heuristic (or even simply a frame of reference) machine learning practitioners generally consider data sets on the order of hundreds of samples to be relatively small [ref 'Sample size planning for classification models', Predicting Sample size required for classification performance, How Much Training Data is Required for Machine Learning?]. The exact size of a data set required to properly train a typical Hidden Markov Model (or any machine learning algorithm in general) depends on a number of different factors including the, method of classification, complexity of the classifier, separation between classes, variance and presence of noise in the data, amongst many other factors. The noisier, the more complex and the greater the variance in the data, typically the larger the dataset required to achieve good performance. There is no upper limit for how much data should be used for training but there point at which increasing input data begins to yield diminishing returns in improving predictive performance [How Much Training Data is Required for Machine Learning?]. The exact relationship between training set size and predictive performance for an algorithm and problem in question is often shown as a 'learning curve' graph (which plots training set size versus prediction error(s)). To the best of the author's knowledge the learning curve for this particular application (or a sufficiently analogous application) has not yet been determined. Given however that we expect that the data collected in this study will be relatively noisy and complex we expect that the model may lean towards requiring more data rather than less data. Therefore, we collected as much data as was available in order to not prematurely limit the power nor the generalizability of the algorithm developed.

2.6 Missing Data

Due to the nature of the data collection methods used we don't expect to have a lot of missing data. However, despite initial expectations missing data is still often an inevitable reality for scientific studies. We handle missing data in the following manner:

Missing Heart Rate and Step Count data is not explicitly reported by the activity tracker. If the device is not being worn it will report both step count and heart rate values as 0; therefore the devices automatically perform imputation of all missing values replacing them with 0.

Blood pressure data which always comes as a set of systolic/diastolic readings are imputed using non-parametric multiple imputation using the 'missForest' R package (which uses a random forest algorithm to predict missing values and makes no assumption about the underlying distribution of the data and supports both missing categorical and continuous variables) [ref: https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/ & https://cran.r-project.org/web/packages/missForest/missForest.pdf]. Weight, symptom questionnaire responses, heart failure treatments to date and prescribed medications are dealt with using the same approach.

Patients without any specified NYHA class (i.e. missing outcome data) are ignored completely, however gaps in NYHA class outcomes are imputed using the same approach as missing predictors (namely blood pressure, weight, etc.).

2.7 Statistical Analysis Methods

Before using the data to train the classification algorithm the predictors will be normalized to values between 0 and 1 (after imputation of missing data as required). The normalization procedure for each of the predictors is as follows:

1. Heart Rate - linear scaling using: $min = 0$, $max = 208 - 0.7 * age$ (where patient's age is measured in years) [ref: http://www.onlinejacc.org/content/37/1/153]
2. Step Count - linear scaling using: $min = 0$, $max = 300$
3. Blood Pressure (Systolic) - linear scaling: $min = 50$, $max = 300$
4. Blood Pressure (Systolic) - linear scaling: $min = 30$, $max = 160$
5. Weight - linear scaling: $min = 0$, $max = 200$ (the maximum capacity of the weight scales used)
6. Symptom Questionnaire Responses -
   (a) Have you fainted?: $no = 0$, $yes = 1$
   (b) Has your Implantable Cardioverter-Defibrillator (ICD) gone off?: $no = 0$, $yes = 1$
   (c) Do you have more chest pain than usual?: $no = 0$, $yes = 1$
   (d) Are you more tired than usual?: $no = 0$, $yes = 1$
   (e) Are you more short of breath than usual?: $no = 0$, $yes = 1$
   (f) Do you feel that your heart is beating unusually?: $no = 0$, $yes = 1$
   (g) Do you feel lightheaded?: $no = 0$, $yes = 1$
   (h) Did you have to stop any of your usual daily activities because of your health?: $no = 0$, $yes = 1$
7. Patient Demographic/Meta Data -
   (a) Sex: $Male = 0$, $Female = 1$
   (b) Age - linear scaling: $min = 18$, $max = 150$
   (c) Height - linear scaling: $min = 50$, $max = 300$
   (d) Handedness: $Left = 0$, $Right = 1$
   (e) Wristband Handedness Preference: $Left = 0$, $Right = 1$
   (f) (Left Ventricle) Ejection Fraction: $min = 0$, $max = 100$
   (g) Heart Failure Diagnosis Date: $min = 0$, $max = 30 * 12$

8. Heart Failure Treatment to Date -
   (a) Lifestyle changes: $no = 0$, $yes = 1$
   (b) Implantable Cardioverter-Defibrillator (ICD): $no = 0$, $yes = 1$
   (c) Left Ventricular Assist Device (LVAD): $no = 0$, $yes = 1$
   (d) Heart transplant: $no = 0$, $yes = 1$
   (e) Other surgical intervention: $no = 0$, $yes = 1$
9. Prescribed Medications -
   (a) I_f Channel Blocker - linear scaling: ivabradine $min = 0$, $max = 15$ [ref: https://www.drugs.com/dosage/ivabradine.html]
   (b) Beta Blockers - linear scaling: Acebutolol $min = 0$, $max = 1200$ [ref: http://reference.medscape.com/drug/sectral-acebutolol-34235 4]
   & Atenolol $min = 0$, $max = 300$ & Bisprolol $min = 0$, $max = 20$ & Carvedilol $min = 0$, $max = 80$ & Labetalol $min = 0$, $max = 2400$ & Metaprolol $min = 0$, $max = 450$ & Nadolol $min = 0$, $max = 320$ & Pindolol $min = 0$, $max = 60$ & Propranolol $min = 0$, $max = 640$ & Timolol $min = 0$, $max = 60$ [Ref: http://www.globalrph.com/beta.htm] [Ref: https://www.heartandstroke.ca/heart/treatments/medications/beta-blockers]
   (c) Aldosterone Antagonists - linear scaling: Spironolactone $min = 0$, $max = 200$ & Eplerenone $min = 0$, $max = 50$ [ref: http://www.globalrph.com/aldosterone_antag.htm] [ref: https://www.heartandstroke.ca/heart/treatments/medications/aldosterone-antagonists]
   (d) Diuretics - linear scaling: Chlorthalidone $min = 0$, $max = 100$ & Ethacrynic acid $min = 0$, $max = 100$ & Furosemide $min = 0$, $max = 600$ & Hydrochlorothiazide $min = 0$, $max = 100$ & Indapamide $min = 0$, $max = 5$ & Metolazone $min = 0$, $max = 20$ [ref: http://www.globalrph.com/diuretics.htm] [ref: https://www.heartandstroke.ca/heart/treatments/medications/diuretics]
   (e) Digoxin - linear scaling: $min = 0$, $max = 15 \times 10^{-3} * weightMax$ [ref: http://www.globalrph.com/antiarrhythmics.htm#digoxin_]

To perform the classification we create and train multivariate hidden Markov models for each of the target risk groups (NYHA class II, III). Since it is probably unreasonable to assume that there no some time-dependence in state changes due to the dynamic nature of human exercise and activity (e.g. people who are performing high-intensity activity are less likely to continue as time goes by since they get tired) we also train equivalent multivariate hidden semi-Markov models for each of the target risk groups to explore the effect of relaxing the Markovian assumption (or time-independence) of pure Markov models.

We identified five candidate sets of variables for inclusion into the predictive models:

1. heart rate + step count
2. the above + weight
3. the above + blood pressure
4. the above + Symptom Questionnaire Responses specifically targeting activity intolerance
5. the above + Prescribed Medications

We opted not to use the 'Heart Failure Treatments to Date' data since we suspected that it would render the model far too complex rendering it even more prone to over fitting and preferred to focus our efforts on the five models already identified.

To further trim the set of variables we propose running a Pearson cross-correlation analysis on the collected data to identify the sets of highly correlated variables. These sets can be trimmed down to the single most useful predictor (whether clinically relevant, easiest to collect, or otherwise). Eliminating needlessly redundant information should help reign in the model complexity.

To verify the classification algorithm we can an internal validation, recognizing that such a validation is likely to produce optimistic results. To verify the algorithm and models we select sets of hidden (semi-)Markov models - one for each risk group - and ask each model in the set to calculate the likelihood that a two-week window of data from a test patient was generated using that model. The model with the highest associated probability predicts the outcome (NYHA class) for that window of test data. By performing this test on each two-week window (incremented by day) for each patient and comparing the predicted value with the most recent physician reported value we can determine the sensitivity and specificity of each classification model set and thereby identify the best model.

2.8 Risk Groups

Patients in the study are classified into one of four risk groups. NYHA class I, II , III & IV. As a specialized tertiary care center the Heart Function Clinic rarely sees NYHA class I patients as they are often asymptomatic with regards to their heart failure, or at least rarely require the specialized level of care offered by the clinic. NYHA class IV patients, while seen at the clinic, often derive little benefit with regards to physical activity monitoring since these patients are known to be often bedridden or severely activity intolerant. Classes II and III however represent a classification challenge since the dividing line between mild and marked activity intolerance can sometimes be unclear especially given the inherent subjective nature of the NYHA classification; training a machine to classify between class II & II patients is therefore likely to produce the greatest benefit to clinical and medical research so we focus on these two risk groups in this study.

NYHA class is typically assessed for every patient with known cardiac disease (usually first objectively verified through the use of some sort of medical imaging modality). It is then reassessed at every clinic visit by the physician responsible for patient's care. At minimum the physician will pose questions to attempt to elucidate the patients' degree of exercise intolerance, for example: "How far can you walk before becoming short of breath?", or "How many flights of stairs can you climb before needing to stop?". The physician will then select an NYHA class based on their clinical experience, professional judgment according to the NYHA class definitions. These definitions are copied below for the reader's convenience [ref: http://professional.heart.org/professional/General/UCM_423811_Classification-of-Functional-Capacity-and-Objective-Assessment.jsp]:

I: "Patients with cardiac disease but without resulting limitation of physical activity. Ordinary physical activity does not cause undue fatigue, palpitation, dyspnea, or anginal pain."

II: "... slight limitation of physical activity. They are comfortable at rest. Ordinary physical activity results in fatigue, palpitation, dyspnea, or anginal pain."

III: "... marked limitation of physical activity. They are comfortable at rest. Less than ordinary activity causes fatigue, palpitation, dyspnea, or anginal pain."

IV: "Patients with cardiac disease resulting in inability to carry on any physical activity without discomfort. Symptoms of heart failure or the anginal syndrome may be present even at rest. If any physical activity is undertaken, discomfort is increased."

2.9 Development vs. Validation

Since we reuse a single dataset the setting, eligibility criteria, outcome, and predictor remain unchanged for both development and (internal) validation.

3 Results

3.1 Participants

3.2 Model Development

3.3 Model Specification

3.4 Model Performance

3.5 Model-updating

4 Discussion

4.1 Limitations

4.2 Interpretation

4.3 Implications

5 Other Information

5.1 Supplementary Information

5.2 Funding

Funding for this project was provided by the:

Centre for Global eHealth Innovation (Heart Function Clinic?)
Canadian Institutes of Health Research (Government of Canada)
Government of Ontario
University of Toronto