# Automated Extraction of the Barthel Index from Clinical Texts

**Phan Giang, PhD[a], Allison Williams, ND, PhD, RN[b] and Lisa Argyros, RN[b]**
**[a]George Mason University, Fairfax, VA; [b]Veteran Affairs Health System, Bay Pines, FL**

**Abstract**

*This paper describes a text mining program that computes the Barthel score of functional status by analyzing clinical notes stored in EHR and comparing them to the textual evidence provided by clinician expert. The program demonstrates high accuracy and overall reliability based on a relatively small number of expert-abstracted charts. It offers an efficient and affordable method for estimating functional status using clinical notes. An important feature of the program is an architecture that facilitates the interaction between users and the program that allows the program to improve its performance based on the feedback from users.*

**Introduction**

Functional status (FS) is a key indicator of patient overall health and quality of life. FS is used to determine care strategies and monitor changes in clinical status. Studies have also shown that FS predicts patient care outcomes. This finding has been demonstrated across diverse patient populations including older adults whose FS is often complicated by frailty, multiple comorbidities, and heterogeneous cognitive abilities. Several instruments are available that capture FS with the more commonly applied being the Barthel Index (BI), Functional Independence Measure, and the Rankin Score (Balu, 2009), (Cohen & Marino, 2000) and (Hobart, et al., 2001). While architecturally different, these tools all capture dimensions associated with ability to perform necessary activities of daily living (ADLs). Indices to measure FS have been developed for research purposes but are not typically an integral part of data collected strictly for clinical purposes. An alternative is to derive the functional measurement scale directly from clinician notes recorded in the patient electronic health records (EHR). In this manner, standardized scores can be reported without requiring additional data collection efforts. These scores can not only inform direct clinical care, but also provide system level information to improve healthcare delivery.

An automated method of assessing patient's functional ability is particularly useful to nursing home administration. Under CMS Nursing Home Quality Initiative, nursing homes are required to collect a Minimum Data Set (MDS) for assessing functional capabilities of each of their residents. The MDS contains quantified ADLs measures that can be cross walked to items on FS indices allowing derivation of a FS score. However, this information is not collected for long-term care patients who do not reside in a Medicare/Medicaid-certified nursing home presenting an information gap for such patients and inability to compare patients across care settings. The need for comparison across long term care settings will become increasingly important as new models of non-institutional long term care delivery emerge. One example of such non-institutional care is the Department of Veterans Affairs Medical Foster Home (MFH) Program. With efforts to evaluate this alternative to nursing home placement, we sought a method for direct FS comparison among veterans in this program compared to veterans in nursing homes.

The MFH program is an innovative housing option in which veterans with long-term care needs reside with a family in the community instead of an institutional setting. The VA initiated the MFH program in 2006 as an alternative to nursing home care. The MFH Program began as a pilot in 2000 in Little Rock, Arkansas. Two more programs followed: one in Tampa, Florida and another in San Juan, Puerto Rico. As of February 2013, the MFH Program is operational and assisting veterans with MFH selection in 86 sites in 42 localities. The program is expanding to 102 sites in 48 states and territories. To date, 1,722 veterans have been served, and there are presently 524 caregivers, 565 Veterans in MFHs nationally.

While there is a general belief that the MFH program is a safe, effective, and preferred model for long-term care, data supporting these assumptions are anecdotal. Additionally, individual variations among caregivers as well as patients that may influence program success are unknown. Given the key role of FS to patient care and outcomes, this variable must be included in any program evaluation. As part of a larger project comparing the safety, efficacy and cost-effectiveness of the MFH program to the CLC, this project develops a tool that scores FS from clinician notes thereby providing a FS measure for MFH patients and reducing data collection burden.

The Barthel Index is applied frequently in geriatric populations to assess functional status according to individual's ability to attend to activities of daily living (ADLS). It captures actual rather than potential capability, and the preferred source of information is direct observation. The instrument contains 10 items: Feeding, Bathing, Grooming, Dressing, Bowels, Bladder, Toilet use, Transfers, Mobility, and Stairs. Each of these items are rated according to level of independence. Figure 1 has an example of the Barthel Index form (Collin, Wade, Davies, & Horne, 1988) with a brief score interpretation. For example, Feeding is rated between three levels: 0 (the patient is unable to feed himself or herself), 1 (needs help cutting, spreading butter, etc.) and 2 (the patient is able to feed himself or herself independently). Items have different weights according to dependence level stratification. For example, Bathing scores is between 0 and 1 while Transferring score ranges from 0 to 3. Scores generally represent ability within the past 24-48 hours; however, a longer or shorter time frame may apply according to clinical status. The sum of itemized scores produces a total score that ranges from 0 to 20 with lower scores indicating lower level of independence. (In literature, the Barthel Index is also measured in the [0, 100] scale. For example, "Feeding" score can be 0, 5 or 10 instead of 0, 1 or 2. Except for the multiplier, there is no difference between the two versions of Barthel Index.) It is noted that changes of more than two points (10%) in total score reliably reflect the genuine changes in FS. Barthel Index has high inter-rater reliability (r = 0.95) and test-retest variability (r = 0.87) as well as high internal consistency (Cronbach's alpha > 0.80) (Cohen & Marino, 2000) (Hobart, et al., 2001). Concurrent validity has been demonstrated through high correlations between the BI and other measures of FS including the Katz Index (k = 0.77) and the Kenny Self-Care Evaluation (Spearman r = 0.73) (Cohen & Marino, 2000).

**Barthel Index of Activities of Daily Living**

***Instructions:*** Choose the scoring point for the statement that most closely corresponds to the patient's current level of ability for each of the following 10 items. Record actual, not potential, functioning. Information can be obtained from the patient's self-report, from a separate party who is familiar with the patient's abilities (such as a relative), or from observation. Refer to the Guidelines section on the following page for detailed information on scoring and interpretation.

**The Barthel Index**

Bowels
0 = incontinent (or needs to be given enemata)
1 = occasional accident (once/week)
2 = continent
*Patient's Score:* ____________

Bladder
0 = incontinent, or catheterized and unable to manage
1 = occasional accident (max. once per 24 hours)
2 = continent (for over 7 days)
*Patient's Score:* ____________

Grooming
0 = needs help with personal care
1 = independent face/hair/teeth/shaving (implements provided)
*Patient's Score:* ____________

Toilet use
0 = dependent
1 = needs some help, but can do something alone
2 = independent (on and off, dressing, wiping)
*Patient's Score:* ____________

Feeding
0 = unable
1 = needs help cutting, spreading butter, etc.
2 = independent (food provided within reach)
*Patient's Score:* ____________

Transfer
0 = unable – no sitting balance
1 = major help (one or two people, physical), can sit
2 = minor help (verbal or physical)
3 = independent
*Patient's Score:* ____________

Mobility
0 = immobile
1 = wheelchair independent, including corners, etc.
2 = walks with help of one person (verbal or physical)
3 = independent (but may use any aid, e.g., stick)
*Patient's Score:* ____________

Dressing
0 = dependent
1 = needs help, but can do about half unaided
2 = independent (including buttons, zips, laces, etc.)
*Patient's Score:* ____________

Stairs
0 = unable
1 = needs help (verbal, physical, carrying aid)
2 = independent up and down
*Patient's Score:* ____________

Bathing
0 = dependent
1 = independent (or in shower)
*Patient's Score:* ____________

***Total Score:*** ____________

(Collin et al., 1988)

***Scoring:***
Sum the patient's scores for each item. Total possible scores range from 0 – 20, with lower scores indicating increased disability. If used to measure improvement after rehabilitation, changes of more than two points in the total score reflect a probable genuine change, and change on one item from fully dependent to independent is also likely to be reliable.

**Figure 1: Barthel Index Form**

Healthcare researchers, administrators and providers become increasingly interested in methods to unlock the valuable stock of narrative information stored in electronic health records (EHRs). Currently, unstructured data in narrative format are underutilized because they are not amenable for automatic processing by computers and the high volume of data makes manual extraction cost prohibitive. In recent years, a number of initiatives to promote the development of Natural Language Processing (NLP) technology for healthcare have been implemented. For example, the National Institute of Health helped to create and provided funding for the Informatics for Integrating Biology and the Bedside (i2b2) with focus on promoting NLP and and text mining applications in healthcare.

## Objective

This study used NLP techniques to extract FS information from clinical notes and computed Barthel scores.

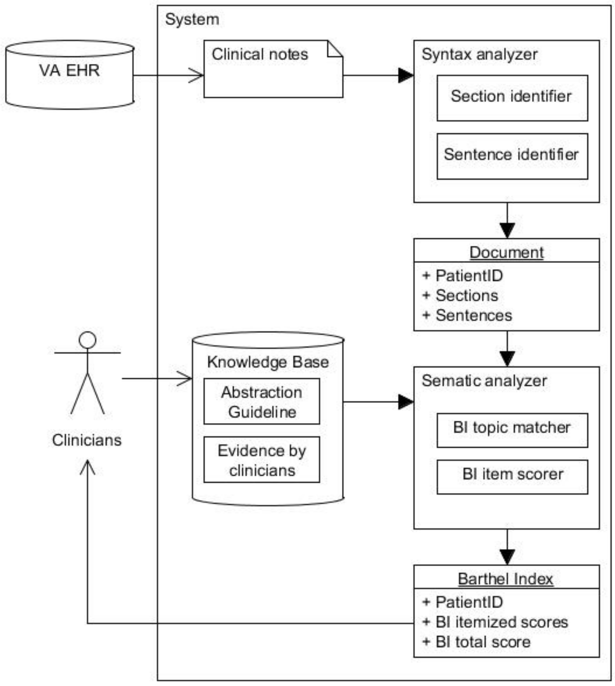


**Figure 2: Architecture of Barthel Index extraction program**

## Method

### *System architecture*

The high-level system architecture of the extraction program is described in Fig. 2. The input for the system consists of clinical notes obtained from the Veterans Health Information Systems and Technology Architecture (VistA) - VA EHR. Each text file (patient chart) consists of tens to hundreds of clinical notes recorded within a pre-determined time period.

The first step of the process is a syntactic analysis that breaks a file into documents and performs various tasks such as sectioning and sentencing based on syntactical rules. As a result, each clinical document is mapped into an object that has basic attributes such as document identification number (ID), document date and also a nested structure of sections, paragraphs and sentences.

The main step of the algorithm is the semantic analysis. Initially, sections known to be irrelevant to the FS extraction such as administrative and demographic information are excluded from further analysis. All other sections are then analyzed sequentially. Within each section, sentences are subjected to a two-step "matching" process. Two pieces of information are evaluated. The first one is the *topic relevancy* and the other is the *functional status level*. For example, a sentence may be relevant to "Feeding" and at the same time provides information that allows inferring the "needs help" level (1). However, it is also possible that the topic relevancy information and FS information are found in different sentences. What remains invariant is that the topic information always precedes the FS information as far as sentences are concerned. This invariance is used in the algorithm to link two pieces of information together.

Steps in the semantic analysis are as follows:

1. First, each sentence is analyzed by the *topic matcher* to decide if it is relevant to one of ten Barthel Index topics. The computed matching score is a number between 0 and 1. If the score is higher than an empirically selected threshold, the sentence is marked as "matched" for the topic.

2. Next, the marking of a sentence triggers the BI item scoring. This applies not just for the marked sentence, but also for the rest of the section in which that sentence is found until a sentence marked with a different topic is encountered.
3. BI item scoring is applied for each FS level according to the item specifications. For example, “Feeding” has three levels, “Bathing” has two while “Transferring” has four levels.
4. If a sentence matched the FS level, a piece of *evidence* is generated that includes the topic, the score and the actual sentence where it is found. The included sentence is used for the diagnostic purpose. Suppose, for example, a clinical note includes sentence “pt. feeds himself with fork and spoon”. The word “feeds” will match the BI “Feeding” topic. Words “feeds” and “himself” will match to a score of 2 which corresponds to an “independent” FS level. The following piece of evidence will be generated (in XML notation): `<evidence><topic>Feeding</topic><score>2</score><support>pt. feeds himself with fork and spoon</support></evidence>`.
5. The BI itemized score for a patient is calculated by averaging the evidence scores for each of ten BI topics. The Barthel (total) score is the sum of itemized scores.

Both components of the semantic analyzer, the topic matcher and the item scorer, use an algorithm that calculates the *semantic similarity* between two sentences. Essentially, in this application, the domain knowledge contains the sentences or phrases that are manually extracted by expert clinicians as the evidence to support their judgment of patient functional status (see below). While scanning clinical notes, the algorithm compares the sentences in the notes with the sentences in the knowledge base and computes their semantic similarity scores. If the similarity score passes a threshold then two sentences are considered “matched”.

Clearly, the goal of the algorithm is to *emulate clinician’s judgment* about patient FS. In other words, the algorithm makes no attempt to “understand” clinical notes, as human experts do, to get the Barthel score. Instead, the algorithm uses a simpler and more efficient approach, that is, to copy the observable behavior of human experts. Specifically, the algorithm tries to pick up the sentences in clinical notes that are similar to the evidence used by experts to arrive at their judgment. Figuratively, the relationship between the human expert and the algorithm is analogous to the relationship between the engineer who designs a machine based on understanding of system requirements and technological properties of materials and the assembly line workers who, following engineer’s instruction, make serial production of the machine much faster.

This approach has a number of implications. First, the accuracy of the algorithm Barthel score cannot exceed the accuracy of clinician judgment about Barthel Index. Suppose that two experts have different judgments or using different pieces of evidences, then that difference will be carried over into the algorithm extraction. Or if some type of evidence is completely missed by human experts, then the algorithm will miss it too.

***Expert provided knowledge***

As illustrated in the Fig. 2, the semantic analyzer uses the knowledge base to analyze clinical notes to create evidences for FS evaluation. Because the domain knowledge provided by clinician experts played a crucial role, in this approach we set out two design criteria for the interaction between the software and its the intended users - the domain experts. First, the interaction must be simple and intuitive. Second, the interaction must be adaptive i.e., the software must be able to improve itself based on the feedback of the users.

In this project, two of the authors with clinical expertise went through a sample of clinical notes, extracted relevant sentences for each of Barthel Index items with FS score. Data for the item “stairs” were infrequently available, hence, this item was omitted. Fig. 3 provides an example of the expert input that is used to create the knowledge base. For example, the presence of sentence “TOILETING/Continent – Continent of elimination” implies that scores for “BOWEL” and “BLADDER” items are 2.

Alternatively, the experts can just go over the clinical notes, directly highlight the relevant sentences and insert their evaluation score. A program will collect the highlighted sentences and put them into the knowledge base.

In addition to the expert extracted sentences, the knowledge base also consists of a glossary of abbreviations, medical terms and the special lexicon that are not found in a standard dictionary of general purpose such as WordNet, a publicly available semantic dictionary of English developed at Princeton University (Fellbaum, 2005) . For example, the glossary of abbreviations found in MFH corpus consist of tokens such as “adls” which stands for “activities of daily living” or “4ww” for “4-wheeled walker”, “wfl” for “within functional limits” etc.

**Bowel**

Continent of bowel and bladder. (BOWEL = 2; BLADDER = 2) Continent of bowels and bladder. (BOWEL = 2; BLADDER = 2)
Continent of B&B (BOWEL = 2; BLADDER = 2)
Continent bowel & bladder (BOWEL = 2; BLADDER = 2) & uses bathroom in room (TOILET = 2)
TOILETING/Continent - Continent of elimination. (BOWEL = 2; BLADDER = 2)

**Bladder**

Continent of bowels and bladder. (BOWEL = 2; BLADDER = 2)
Continent of B&B (BOWEL = 2; BLADDER = 2)
Continent bowel & bladder (BOWEL = 2; BLADDER = 2) & uses bathroom in room (TOILET = 2)
Continent of urine, (BLADDER = 2)
TOILETING/Continent - Continent of elimination. (BOWEL = 2; BLADDER = 2)
GU:continent of urine (BLADDER = 2),
Continent of bowels and bladder. (BOWEL = 2; BLADDER = 2)
GU:
BIOPHYSICAL ASSESSMENT
WNL: Able to void without pain or difficulty, continent, urine clear, yellow.

**Grooming**

He was assisted with set up of oral care and then he brushed own teeth. (GROOMING = 1)
Grooming: Extensive assistance (GROOMING = 0)
DRESSING/GROOMING/BATHING - Independent with minimal assistance with bathing. (DRESSING = 2; GROOMING = 1; BATHING = 1)

**Figure 3: Sample of expert provided input**

In this application, the expert knowledge in the form of extracted sentences can be used in two different modes. Originally, it is used as the seed knowledge to train the system. However, after the initial training, we allow human users to use the program in an interactive mode. Clinicians can improve the performance of the program by adding more sentences it misses and rejecting the false positive evidence that the program captures. This interactive feature enables the program to improves itself without the intervention of the software developers.

***Semantic matching***

The semantic matching in this application is an algorithm based on word occurrence and the semantic relationship between words.

Given the body of expert evidence, for each word in the collection, we compute numbers called the *relevancy weighst* by a formula similar to the popular TF-IDF or the Term Frequency - Inverse Document Frequency technique (see for example (Hand, Mannila, & Smyth, 2001)). The term frequency component measures how often a given word appeared in the expert evidence related to a topic or a FS level. The inverse document frequency component measures how uncommon the word is in the general corpus, not just in the relevant sentences but also in irrelevant sentences. For example, for the "Bowel" topic, words such as "continent", "bowel", "toilet" have high relevance weights while other auxiliary words have low relevance weights. There are different relevance numbers. The topic relevancy is used for the topic matching and the FS relevancy is used in the BI item scorer.

Suppose that $S_1$ is a sentence in the domain knowledge base and $S_2$ is another sentence in a clinical note currently under examination. The similarity of $S_2$ to $S_1$ reflects the extent to which weighted words in $S_1$ are "found" in $S_2$. The meaning of "found" in our algorithm is understood quite broadly. Not only are literal occurrences of a word counted as "found", but also synonyms and various grammatical forms. To account for the ambiguity, the matching between two words is graded between 0 and 1 where 1 means a exact occurrence and 0 means completely unrelated. The list of synonyms and grammatical forms of a word is obtained by using MIT Java WordNet Interface (Finlayson, 2011).

**Data**

The set of 30 patient charts, manually abstracted by two experts, arrived in two batches at different times. The first batch has 10 patients. The evidence for those patients is used to *train* the program (instantiate the knowledge base). The later batch of 20 patients is used to *test* the performance of the program. The size of each patient chart, consisting of tens to hundreds clinical notes, ranges from 200K to 2,000K, or equivalently, 5,000 to 50,000 text lines. On average, it took an expert about 4 hours to manually complete one chart abstraction. All of 30 patient charts are selected randomly from the Medical Foster Home (MFH) patient pool. The representativeness of the sample is important because if sentences or phrases of a specific type were systematically missed from the expert knowledge input then the algorithm would also fail to capture them in clinical notes.

## Results

The algorithm was implemented in Java. Running on Windows 7 system with Intel Core5 processor, it takes on average about 2.5 minutes to complete scoring of a patient chart. We have tested the algorithm performance on two data sets.

**Test case 1: 20 MFH patients.** Training data includes 10 manually abstracted patient charts from MFH pool. The test data includes another sample of 20 patient charts from the same MFH pool that is randomly selected for manual abstraction.

The manually abstracted Barthel scores are plotted against the scores calculated by the program (see Fig. 3). The Mean Square Error MSE (Algorithm Barthel score - Expert Barthel score)$^2$ is 7.14. The Root Mean Square Error (RMSE) is 2.67 or 13.35% of the score range (20). This number can also be interpreted relatively to the reliability of Barthel Index as a measure of FS. Studies have shown that changes of more than two points in the total Barthel score reflect a probable genuine change in FS after rehabilitation (Collin, Wade, Davies, & Horne, 1988). Thus, the reliability of the algorithm score with respect to expert Barthel score is comparable to the reliability of expert Barthel score with respect to the true FS.

The relationship between the expert score on the algorithm score can be understood via standard linear regression in which the expert score is treated as the dependent variable and the algorithm score as the independent variable. Using this regression equation, one can estimate the unobserved expert score from a given algorithm score. A perfect relationship between the two scorings (algorithmic and expert) would be described by a regression line with the zero intercept and the unit slope. The closer the slope to the unity the better and the closer to zero the intercept the better. The actual regression shows a small and negative intercept (-0.95) and the unity slope (1.0021) which together with high adjusted R-Squared value of 0.789 indicate a very good agreement between the algorithm scores and the gold standard of expert scores.

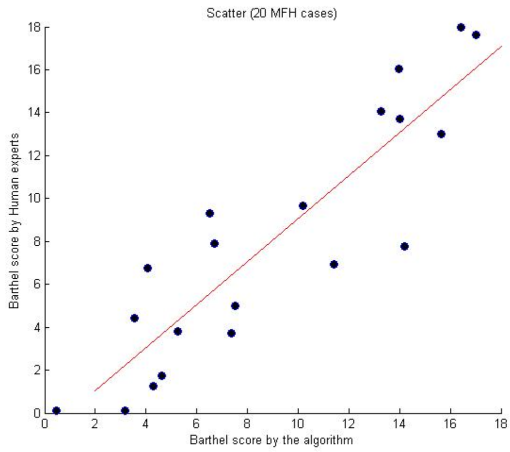


**Figure 4: Expert abstracted Barthel score vs. Algorithm Barthel score (MFH corpus)**

***The linear regression output***

```
Linear regression model:
    y ~ 1 + x1
Estimated Coefficients:
                   Estimate     SE          tStat       pValue
    (Intercept)    -0.95622      1.2148    -0.78715       0.44143
    x1               1.0021    0.11823      8.4765    1.0647e-07
Number of observations: 20, Error degrees of freedom: 18
Root Mean Squared Error: 2.64
R-squared: 0.8,  Adjusted R-Squared 0.789
F-statistic vs. constant model: 71.9, p-value = 1.06e-07
```

Measure such as Precision, Recall and their harmonic average, the F1-measure, are routinely used to evaluate the performance of text mining and information retrieval (Uzuner, Solti, & Cadag, 2010). In this application, because the variable of interest, Barthel score, is quantitative rather than dichotomous (True/False) those measures are not very helpful. However, to facilitate a comparison of our performance numbers with that of other methods in literature, we create binary class labels for patient charts based on the expert Barthel score. We define "*High functional*" as "expert Barthel score is higher than 10" and "*Low functional*" if "expert Barthel score is less than or equal to 10**"**. Clearly, this conversion loses some information because, for example, it would classify two patients with Barthel scores 11 and 20 respectively with the same label "High functional" even they are fairly different in terms of FS. Acknowledging the limitation, we calculate ROC curve and the AUC (0.9643) parameters that provide more information than a Precision-Recall pair.

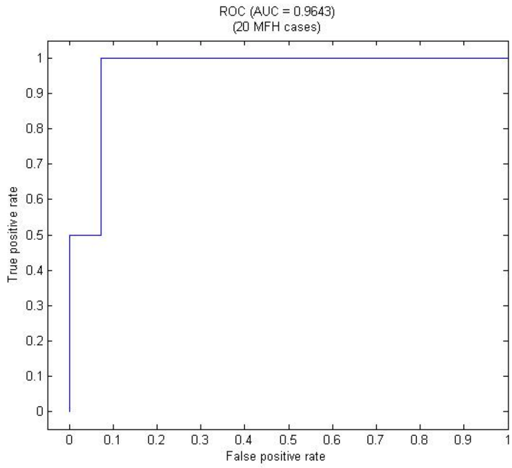


**Figure 5: ROC curve and AUC (MFH corpus)**

**Test case 2: 123 CLC patients**. Previously, the program was tested on 20 MFH patient charts after training on a collection of 10 MFH charts. In this test, the program was run on a set of 123 patients charts randomly selected from the pool of patients who reside in VA Nursing Homes known as Community Living Centers (CLC). These charts include Minimum Data Set (MDS) data required by the Medicare Nursing Home Quality Initiative. Barthel scores for CLC patients can be derived from MDS data. We want to know how the algorithm, using the sentences extracted by experts from MFH charts, predicts the MDS Barthel scores available for veterans residing in CLCs.

The training data include all 30 expert abstracted charts for MHF patients. The testing data are 123 CLC patient charts for whom MDS Barthel scores are available. The MSE is 9.21 (RMSE is 3.04). Linear regression of the MDS Barthel score against the algorithm score is available in Fig. 6. The intercept is 2.0518 and the slope is 0.81155. The values of R-squared and Adjusted R-Squared are 0.542 and 0.538 respectively.

## Discussion

Comparing performance of two tests, we see some decline in performance when the algorithm is applied to a text corpus from which it does not have training data. The deterioration of performance on CLC corpus compared with MFH corpus can be seen in terms of RMSE (3.04 vs. 2.67) and the calculated parameters of the regression lines: intercept (2.05 vs. -0.95) and slope (0.81 vs. 1.0). The effect can also be seen in terms of AUC (0.96 vs. 0.88) and ROC curve. Nevertheless, under this worst case scenario, the performance numbers are still practically useful and are comparable to results reported in medical text mining literature (see the discussion of related works). For example, along the ROC curve, at a false positive rate (FPR) of 0.2, the algorithm attains true positive rate (TPR) of 0.8; at FPR of 0.4, the TPR is 0.9. Overall AUC value of 0.88 shows the robustness of the algorithm.

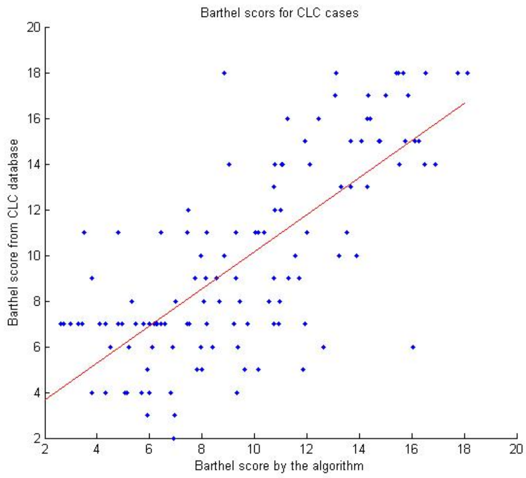


**Figure 6: MDS Barthel score vs. Algorithm Barthel score (CLC corpus)**

### ***The linear regression output***

```
Linear regression model:
    y ~ 1 + x1
Estimated Coefficients:
                   Estimate    SE          tStat      pValue
    (Intercept)     2.0518       0.7099    2.8903      0.0045628
    x1             0.81155     0.067865    11.958     3.0989e-22
Number of observations: 123, Error degrees of freedom: 121
Root Mean Squared Error: 2.96
R-squared: 0.542,  Adjusted R-Squared 0.538
F-statistic vs. constant model: 143, p-value = 3.1e-22
```

Under the same convention defined earlier for “High functional” and “Low functional”, the ROC curve and AUC (0.877) are calculated.

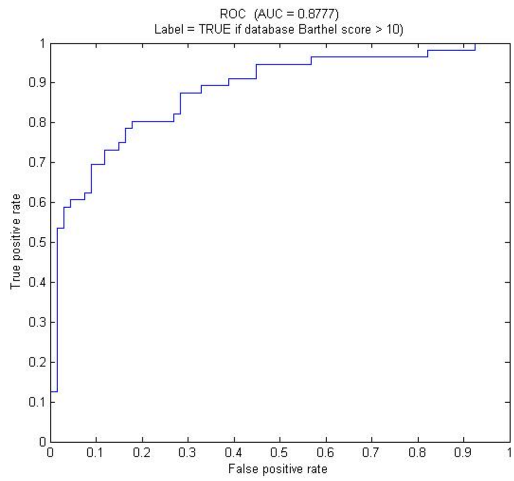


**Figure 7: ROC curve and AUC (CLC corpus)**

Some decline in performance is not surprising. There are at least two plausible factors which are supported by our ex-post diagnostic analysis. The most obvious one is the mismatch in terms of textual and linguistic characteristics between the expert sentences provided for training which were taken from the MFH corpus and the clinical notes that the algorithm analyzes (the CLC corpus). Because the algorithm works by computing the semantic similarity between the training sentences and that in the clinical notes, relevant sentences from the same corpus have a greater chance of being picked up. Put it differently, some semantically relevant sentences in CLC notes could be missed because there are no similar sentences in the knowledge base. To address this issue, it is necessary to add into the knowledge base the sentences extracted by experts from CLC corpus.

The second factor that contributes to the decline is the mismatch in terms of scope between the body of clinical notes that is used by the algorithm to compute Barthel score and the score inferred from MDS data. Typically, the CLC clinical notes fed into the algorithm reflect the patient health situation over a relatively long period (about six months) while the Barthel scores inferred from MDS is just a snapshot (24-48 hours). During the longer period, patient's health condition may change significantly. For example "Feeding" FS may regress from "independent" to "needs some help". A snapshot view like MDS likely can capture only one of the states while the clinical notes have evidence for both FS levels. The act of averaging out all the evidence would lead to a score that is different from the snapshot score. This conclusion is supported by an ext-post manual analysis of several outlier cases where the difference between algorithm score and MDS score is large. One of such cases represented by a point in the plot (Fig. 6) where algorithm score is 16 and MDS Barthel score of 6. This problem could be addressed by having a narrower window for the clinical notes to be used as input of the algorithm.

## Related works

Various machine learning techniques have been tried for clinical text mining problems. The most commonly used methods are rule-based inference and statistical classification models such as logistic regression, support vector machine and its variants. For example, (McCart, Berndt, Jarman, Finch, & Luther, 2012) wanted to identify falls within clinical text associated with an ambulatory encounter. Their data consisted of 26,010 annotated documents from 2,241 patients who were selected based on fall-related ICD-9-CM E-code. They trained three different statistical text mining using logistic regression and support vector machine (SVM) models on 70% of the data tested the models on the remaining 30% of documents. They reported high AUC scores of 95% or more for all three methods. (Botsis, Nguyen, Woo, Markatou, & Ball, 2011) used 6034 US Vaccine Adverse Event Reporting System (VAERS) reports for H1N1 vaccine with positive/negative for anaphylaxis classification by medical experts for training. They showed that rule-based classifier and classifiers based on boosted trees and SVM had good performance in terms of Recall, but suffered high misclassification rates. Patrick and Li (Patrick & Li, 2010) used a combination of two machine learning algorithms and several rule-based engines to extract medication information in the 2009 i2b2 information extraction challenge. They reported good results with approximately 90% accuracy on five out of seven entities in the name entity recognition task, and an F-measure greater than 95% on the relationship classification task. Gundlapalli et al (Gundlapalli, 2012) extracted surveillance information about homelessness status of patients from the clinical notes in the VA EHR. They showed that the templates existing within the text require special treatment, and including these templates into analysis improves the accuracy. Kraus et al (Kraus, Blake, & West, 2007) compared several methods used to identify drug, dosage, and method of delivery information from transcribed physician notes. They showed that using just one extraction heuristic can achieve an average precision 96.70% and 79.72% recall. They argued that a simple method using a small number of heuristics can provide accurate extraction of drug, dosage and method of delivery information in medical notes. Deleger et al (Deleger, Grouin, & Zweigenbaum, 2010) described another approach using a semantic lexicon and extraction rules to extract medication information from clinical records in the context of the i2b2 2009 challenge with performance of 77% (F-measure).

Our algorithm differs from those text mining applications in two important aspects. First, our program is not a binary classifier because the variable of interest is quantitative not dichotomous. The training data provided by experts do not have binary labels either. Second, arguably more important distinction is our approach to expert domain knowledge. The main goal of this text mining program is to emulate the evaluation of the clinicians. Unlike other applications that demand large amount of manually labeled or annotated data for training, this algorithm can run with relatively small numbers of expert-abstracted charts. Of course, more training data is always better but high initial demand for training data is an obstacle in practice.

## Limitations and future works

A number of known issues that have not been addressed in this version of the application due to limited data availability and resource limitations. For example, the temporal information in the clinical notes is mostly ignored because the notes, before they can be fed to the algorithm, have undergone a de-identification process which, among other things, scrambles the factually recorded dates. Another issue is the difficulty in recognizing the intention of the sentences. The narrative clinical texts contain not only factual observations by clinicians but they also include “formulaic” templates and forms which do not represent factual information. At this point, the program is not able to differentiate the intention behind the texts. Finally, we recognize the fact that the body of expert provided evidence used to instantiate the knowledge base is still small and exclusively from the MFH cohort. This explains the deterioration when applied to veterans residing in CLCs. However, improvement in performance can be achieved if more domain knowledge input is available.

## Conclusion

This paper describes a text mining program that computes the Barthel score of functional status based on analyzing clinical notes and comparing them with expert provided textual evidence. The initial results showed accuracy and reliability based on a *relatively small number* of expert-abstracted charts. The Barthel score computed by the program is strongly correlated with the gold standard, expert-abstracted Barthel scores. The main advantage of this text mining approach is the ability to take advantage of the vast amount of clinical notes based on relatively modest demand on training data. The program offers a much more efficient and affordable alternative to manual extraction. This application is different than other clinical text mining applications in several aspects. We extract a quantitative variable from the narrative texts rather than binary classification labels. An important feature is an architecture that facilitates the interaction between users and the program that allows the program to improve its performance based on the feedback from users.

## Reference


Balu, S. (2009). Differences in psychometric properties, cut-off scores, and outcomes between the Barthel Index and Modified Rankin Scales in pharmacotherapy-based stroke trials: Systematic literature review. *Current Medical Research and Opinions , 25* (6), 1329-134.

Botsis, T., Nguyen, M. D., Woo, E. J., Markatou, M., & Ball, R. (2011). Text mining for the Vaccine Adverse Event Reporting System: medical text classification using informative feature selection. *J Am Med Inform Assoc ,* 631-638.

Cohen, M., & Marino, R. J. (2000). The tools of disability outcomes research functional status measures. *Archives of Physical Medicine and Rehabilitation , 81 (suppl 2)* , S21-S29.

Collin, C., Wade, D. T., Davies, S., & Horne, V. (1988). The Barthel ADL Index: a reliability study. *International disability studies , 10* (2), 61-63.

Deleger, L., Grouin, C., & Zweigenbaum, P. (2010). Extracting medical information from narrative patient records: the case of medication-related information. *J Am Med Inform Assoc*, 17-55.

Fellbaum, C. (2005). WordNet and wordnets. In K. Brown, *Encyclopedia of Language and Linguistics,* pp. 665-670.

Finlayson, M. (2011). *The MIT Java Wordnet Interface.* Cambridge, MA: (http://projects.csail.mit.edu/jwi/).

Gundlapalli, A. (2012). Extracting Surveillance Data from Templated Sections of an Electronic Medical Note: Challenges and Opportunities. *International Society for Disease Surveilance 2012 Conference Abstracts.*

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of data mining.* MIT Press.

Hobart, J. C., Lamping, D. H., Freeman, J. A., Langdon, D. W., McLellan, D. L., Greenwood, R. J., et al. (2001). Evidence-based measurement: Which disability scale for neurologic rehabilitation? *Neurology , 57,* 639-644.

Kraus, S., Blake, C., & West, S. L. (2007). Information Extraction from Medical Notes. *MEDINFO.*

McCart, J. A., Berndt, D. J., Jarman, J., Finch, D. K., & Luther, S. L. (2012). Finding falls in ambulatory care clinical documents using statistical text mining. *J Am Med Inform Assoc doi:10.1136/amiajnl-2012-001334* .

Patrick, J., & Li, M. (2010). High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *J Am Med Inform Assoc* , 17-52.

Uzuner, Ö., Solti, I., & Cadag, E. (2010). Extracting Medication Information from Clinical Text. *J Am Med Inform Assoc , 17,* 514-518.