Regular Expression-Based Learning for METs Value Extraction

Douglas Redd, MS [1,2], Jinqiu Kuang, MS [1], April Mohanty, PhD [1], Bruce E. Bray, MD [2], Qing Zeng-Treitler, PhD [1,2]

[1]VA Salt Lake City Health Care System; [2]Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah

Abstract

Functional status as measured by exercise capacity is an important clinical variable in the care of patients with cardiovascular diseases. Exercise capacity is commonly reported in terms of Metabolic Equivalents (METs). In the medical records, METs can often be found in a variety of clinical notes. To extract METs values, we adapted a machine-learning algorithm called REDEx to automatically generate regular expressions. Trained and tested on a set of 2701 manually annotated text snippets (i.e. short pieces of text), the regular expressions were able to achieve good accuracy and F-measure of 0.89 and 0.86. This extraction tool will allow us to process the notes of millions of cardiovascular patients and extract METs value for use by researchers and clinicians.

Introduction

Functional status (FS) is an important clinical outcome and has been referred to as the sixth vital sign[1]. In the treatment of cardiopulmonary diseases, improving patients' FS, particularly cardiopulmonary capacity, is a high priority.

Atrial fibrillation (AF) is a common disturbance of the heart's electrical conduction system that causes an irregular heart rate. Impaired FS is often a symptom of AF, including reduced exercise tolerance in 15-20% of patients[2, 3]. Therefore, one AF treatment goal is to improve FS. FS measurement can help guide treatment strategies and facilitate examination of treatment response[4].

Maximal oxygen consumption is considered the best metric to evaluate exercise capacity, which is a reflection of FS. Patients having poor FS due to cardiopulmonary disease, or other factors, have low maximal oxygen consumption. Clinically, exercise capacity in the form of Metabolic Equivalents (METs) can be measured in several ways (treadmill tests, bicycle ergometry, six minute walk tests, etc.). One MET is defined as the amount of oxygen consumed while sitting at rest, two METs as twice the amount of oxygen consumed while sitting at rest, etc[5]. Functional Capacity, expressed in METs, has been classified in the perioperative risk literature in several different ways[6, 7]. The Goldman classification (<2 METs: class IV, 2-<5 METs: class III, 5-7 METs: class II, and >7 METs: class I)[7] is commonly selected for studies because it correlates to the New York Heart Association (NYHA)[8] symptom classification criteria. FS, as measured by exercise capacity, is important for quality and life, and also longevity.

Several studies in different populations demonstrate a graded relationship between decreasing METs and increasing mortality[9]. This relationship was specifically studied in the U.S. Department of Veterans Affairs (VA) population. Kokkinos et al. found that patients who could achieve ≥9 METS had a 61% lower hazard for death compared to those with poor exercise capacity (≤4 METS)[9].

While most AF patients have been tested for exercise capacity, the METs values are commonly embedded in free text reports. To make this clinical variable available for health services and clinical research, we developed and tested an NLP module using an algorithm called REDEx (regular expression discovery extraction). Previous work in learning of regular expressions has suffered from problems of limited generalizability and adaptability[10]. REDEx, a supervised learning algorithm that learns regular expressions for value extraction, is designed to address those limitations.

Materials and Methods

Overview

METs is a continuous numerical variable which can be interpreted in a categorical fashion. METs may be reported as a single value (e.g. 5.6), a range (e.g. <4, or 4 to 5), a value with a confidence interval (e.g. 7(+/-)1), or a fuzzy value (~ 4). We also need to filter false positives (e.g. 5 METS sometimes meant 5 metastases, predicted METs need to be filtered but not estimated METs).

Data was taken from the collection of veterans' medical records from the Veterans Health Administration (VHA), aggregated from the multiple VistA electronic medical record (EMR) systems. The collection is maintained by the Veterans Informatics and Computing Infrastructure (VINCI)[11], which is an effort of the U.S. Department of Veterans Affairs to make data available for approved research, and is a collaboration between the VA Office of Information and Technology and the VA Office of Research and Development. Data used in this study was derived from medical records in VINCI through fiscal year 2013. All data access was approved through the local IRB and VINCI policies.

When an exercise stress test (ETT) is performed within the VHA system, results are usually available in a specific type of ETT report. Because VHA is nationwide, the ETT report formats differ between sites and are often templated. Users of VHA include those with dual use of VHA and non-VHA healthcare services; When ETT is performed outside VHA, METs is often described in other types of narrative text such as cardiology reports and outpatient notes. In VINCI, METs can only be found in these free text notes.

To develop an NLP module for METs extraction, we first extracted a set of text snippets using expert-provided keywords. The experts were cardiac clinicians having familiarity with the VHA system. The snippets were then annotated manually. Finally, we adapted the REDEx algorithm which was created by us in a prior study for the extraction of weight values and units.

Data Set and Annotation

The data set was constructed using VINCI in multiple phases: document retrieval, snippet extraction, and annotation:

Document Retrieval: Documents were retrieved using the Voogo search tool for the wider purpose of finding functional status measures, which includes METs measurements[12]. The VINCI database was queried for patients diagnosed with AF or atrial flutter (determined by presence of International Classification of Diseases, Ninth Revision (ICD-9) codes 427.31 or 427.32), and having clinical notes containing keywords "Bruce", "Naughton", "METs", or "VO2". The Bruce protocol and Naughton protocol are common ETT protocols, with outcomes commonly measured in METs and oxygen volume uptake (VO2). Documents containing the keywords were only retrieved when the patient also met the diagnosis criteria.

Snippet Extraction: Snippets of text were extracted from the matching clinical notes, consisting of the matched keywords and their contexts. Each snippet consisted of the keyword, the 20 words preceding the keyword, and 20 words following the keyword.

Annotation: METs values were annotated in the snippets using the Visual Tagging Tool (VTT) originally developed at the U.S. National Library of Medicine (NLM)[13]. Only values were annotated, not the keywords themselves. METs values were represented in various ways, including distinct values and ranges of values (e.g. "5", "5-7", "< 4", "greater than 5"). Many instances were found where embedded forms, or templates, were used to indicate ranges of values. In these cases, multiple ranges were listed but only one was checked (e.g. "[ ] < 3; [x] 3-5; [ ] 5-7; [ ] > 7"). In these instances, only the checked value was annotated. Examples are shown in Table 1.

REDEx

A REDEx model was trained using the annotated snippets and validated using 10-fold cross validation. We developed the REDEx algorithm previously to discover regular expressions for extracting values for body weight and weight units[14]. We adapted the REDEx algorithm for METs value extraction by (1) converting tokens to more general regular expressions where possible (table 2, step 3); (2) generalizing the capture group by aggregating different labeled segments (table 2, step 4); (3) performing two-tiered discovery. REDEx first generates the most generalizable regular expressions for each snippet through a series of steps of abstraction and testing. All expressions are then synthesized to reduce redundancy. This gives the resulting regular expressions more generalizable applicability. In the two-tiered discovery, the first tier uses a test function that does not allow any false positives, which results in a set of highly precise regular expressions. The second tier allows regular expressions to result in false positives as long as there are a higher number of true positives, resulting in a set of regular expressions with better recall. During extraction, matches are taken from the first tier if possible. If no matches are found using the first tier, the second tier of regular expressions is applied. Pseudo-code for the modified algorithm is given in Figure 1. The REDEx pseudo-code is applied twice, once for tier 1 and once for tier 2, with only the test function differing between the runs.

An example of the processing of a single snippet is shown in Table 2. Step 1 shows an example where any length of whitespace characters (e.g. spaces, tabs, carriage returns, line feeds) is replaced with a more general regular expression matching any whitespace up to 50 characters long (\s{1,50}). Also in step 1, any length of digits is replaced with a more general regular expression that recognizes any digits (\d+). Step 2 shows the example after tokens were trimmed from the front and back until causing the test function to fail. Step 3 of the example shows a word that was replaced with a more general regular expression matching any word of similar length (\S{1,3}). The other words could not be replaced without causing the test function to fail. Step 4 of the example demonstrates the replacement of the original labeled segment capture group (\d\.\d+) with a regular expression to recognize all of the labeled segments of all of the snippets.

Training/Testing

A reference standard was created from 2701 annotated snippets, such that each snippet had at most one value for METs. The REDEx model was trained using this reference standard and evaluation was performed using 10-fold cross validation. Each snippet was scored as: true positive if a value was extracted and the value matched the reference value; false positive is a value was extracted but it did not match the reference value (or there was no reference value); false negative if no value was extracted but there was a reference value; or true negative if no value was extracted and there was no reference value.

Table 1 Examples of METs occurrences in clinical notes, with the correct annotated values.

Sample	Annotated Value
estimated peak oxygen consumption was 10 METS.	10
Diagnosis: colon cancer with mets.	(none)
BRUCE: 5 METS 67% MPHR achieved.	5
ETT: Fair (4-6 METS)	4-6
EXERCISE TOLERANCE: > 4 METS, unlimited lifestyle	> 4
obtained fair exercise workload (greater than 5 mets)	greater than 5
Functionality capacity is: [x] <4 Mets [ ] 4-9 METS [ ]10 METs or greater.	<4
Exercise tolerance:doesn't appear to meet 4 METS	(none)
submaximal 11 Mets, HR 140, isolated PVC's	11
5___Lowest Mets with positive ST criteria	5
unsable to exercise > 4 METS per hsitory,	(none)
Min. MPH % METS HR BP RPE Comments 0--2 1.5 0 2.1 94 1 2--6 2.0 0 2.5 98 119/59 3 6--10 2.0 2	2.1, 2.5
He has a METS less than 3.	less than 3

```
BLS = Before Labeled Segment
LS = Labeled Segment
ALS = After Labeled Segment
PS = Positive Text Snippets
NS = Negative Text Snippets
RS = Result set of regular expressions
CG = Capture group
testf = test function. For tier 1 it is match (NS, p’’). For tier 2 it is |match(PS, p’’)|>|match(PS, p’’)|

Regular Expression Discovery (PS, NS)
RS={};
CG={};                                      /* Initialize Result */

for each p in PS {                          /* For each positive text snippet */
    p’=generalization(p);                   /* Replace numbers, whitespace, and punct. with reg. expressions */
    if match (NS, p’)                       /* Test for false positives */
        RS = RS + p;                        /* Add non-generalizable snippets to result set */
    else
        trim = true;
        while trim = true & length (ALS, BLS) > 0   /* Trim generalized expressions */
            p’’ = trim(p’, ALS, BLS);       /* Iteratively remove a token from ALS or BLS */
            if testf (PS, NS, p’’)          /* Apply test function */
                RS = RS + p’;               /* Revert trimming and add to result */
                trim = false;
            else
                p’ = p’’;                   /* Accept trimming and continue */
                subst = true;
            end if
        end while
        subst = true;
        while subst = true                  /* Iterate over terms, least frequent to most frequent */
            p’’ = subst(p’, ALS, BLS);      /* Substitute regular expressions for terms */
            if testf (PS, NS, p’’)          /* Apply test function */
                RS = RS + p’;               /* Revert substitution and add to result */
                subst = false;
            else
                p’ = p’’;                   /* Accept substitution and continue */
                subst = true;
            end if
        end while
        for each c in LS                    /* Iterate over all labeled segments */
            c = generalization(c)           /* Replace numbers with regular expressions */
            CG = CG + c;                    /* Add the labeled segment to the capture group set */
        end for
        for each r in RS
            replace(c, CG)                  /* Replace the capture group with the full capture group set */
        end for
    end if
end for
```

Figure 1 Pseudo-code describing the extended REDEx algorithm for regular expression discovery for extraction. The pseudo-code is applied twice with different test functions (*testf*) for two tiers of regular expressions.

Table 2 Example of regular expression discovery steps for a single snippet. Regular expressions from all snippets are synthesized to reduce redundancy.

<table>
<tr><td>Original Snippet</td><td>LLSB
Abd: soft, nt, nd, nabs
Ext: warm, trace edema

Recent labs reviewed.

ETT-SPECT, XX/XX/XX:
The patient exercised on a treadmill for 7 minutes at a maximum workload of (7.4) METs before stopping because of fatigue. The peak heart rate and blood pressure were 120 beats per minute (79 % MPHR) and 149/65 mm.... The entire left ventricle is well perfused at</td></tr>
<tr><td>Step 1: Generalized numbers and whitespace</td><td>LLSB\s{1,50}Abd:\s{1,50}soft,\s{1,50}nt,\s{1,50}nd,\s{1,50}nabs\s{1,50}Ext:\s{1,50}warm,\s{1,50}trace\s{1,50}edema\s{1,50}Recent\s{1,50}labs\s{1,50}reviewed\.\s{1,50}ETT\-SPECT,\s{1,50}\d+/\d+:\s{1,50}The\s{1,50}patient\s{1,50}exercised\s{1,50}on\s{1,50}a\s{1,50}treadmill\s{1,50}for\s{1,50}\d+minutes\s{1,50}at\s{1,50}a\s{1,50}maximum\s{1,50}workload\s{1,50}of\s{1,50}(\d+\.\d+)\s{1,50}METs\s{1,50}before\s{1,50}stopping\s{1,50}because\s{1,50}of\s{1,50}fatigue\.\s{1,50}The\s{1,50}peak\s{1,50}heart\s{1,50}rate\s{1,50}and\s{1,50}blood\s{1,50}pressure\s{1,50}were\s{1,50}\d+\s{1,50}beats\s{1,50}per\s{1,50}minute\s{1,50}\(\d+\s{1,50}%\s{1,50}MPHR\)\s{1,50}and\s{1,50}\d+/\d+\s{1,50}mm\.\.\.\.\s{1,50}the\s{1,50}entire\s{1,50}left\s{1,50}ventricle\s{1,50}is\s{1,50}well\s{1,50}perfused\s{1,50}at</td></tr>
<tr><td>Step 2: Trimmed until test function failed</td><td>of\s{1,50}(\d\.\d+)\s{1,50}METs</td></tr>
<tr><td>Step 3: Generalize tokens</td><td>\S{1,3}\s{1,50}(\d\.\d+)\s{1,50}METs</td></tr>
<tr><td>Step 4: Add variants of capture group</td><td>\S{1,3}\s{1,50}(between\s{1,50}\d+\s{1,50}and\s{1,50}\d+|\d+\s{1,50}\d+/\d+|greater\s{1,50}than\s{1,50}\d+|>>\s{1,50}\d+|\d+\+|\d+\.\d+\+/\-\d+\.\d+|\d+\-\d+|exceeds\s{1,50}\d+|\d+,\d+|\+\d+|>\s{1,50}\d+|\d+\.\d+\s{1,50}\-\s{1,50}\d+|Normal\s{1,50}to\s{1,50}\d+\.\d+|four|\d+:\d+|\d+\.\d+\s{1,50}to\s{1,50}\d+\.\d+|>\d+\-\d+|over\s{1,50}\d+|greater\s{1,50}than\s{1,50}or\s{1,50}equal\s{1,50}to\s{1,50}\d+|\d+\s{1,50}\-\s{1,50}\d+\.\d+|est\s{1,50}\d+|\d+\.\d+\-\d+\.\d+|~\d+|\d+\s{1,50}\-\s{1,50}\d+|\d+\s{1,50}to\s{1,50}\d+|\d+|exceeding\s{1,50}\d+|\d+<\s{1,50}but\s{1,50}<\s{1,50}\d+|LESS\s{1,50}THAN\s{1,50}\d+|<\s{1,50}\d+|at\s{1,50}least\s{1,50}\d+|\d+\.\d+|less\s{1,50}than\s{1,50}\d+|Two|Less\s{1,50}than\s{1,50}\d+|hardly\s{1,50}\d+|\d+\+\s{1,50}or\s{1,50}minus\s{1,50}one|\d+\-\d+\.\d+|\d+\.\d+\-\s{1,50}\d+\.\d+|MORE\s{1,50}THAN\s{1,50}\d+|\d+\s{1,50}or\s{1,50}Less|>=\d+\.\d+|>\d+|~\s{1,50}\d+|>\s{1,50}\d+\.\d+|<\d+|>\s{1,50}or\s{1,50}=\s{1,50}\d+)\s{1,50}METs</td></tr>
</table>

Results

We performed a 10-fold cross validation of the REDEx model using the 2701 manually annotated snippets, with results summarized in Table 3. We aggregated the results of each of the 10 folds to arrive at the final results. Precision was high at 93% and recall was acceptable at 81%. The F1-score of 86% and accuracy of 89% were very good. The number of true negatives was approximately 1.5 times greater than the number of true positives, which was acceptable.

Table 3 Confusion matrix showing accuracy of the trained REDEx model.

	Reference Standard			
REDEx	TP	855	FP	63
	FN	206	TN	1377

Precision	0.9314
Recall	0.8058
Specificity	0.9562
F1-score	0.8641
Accuracy	0.8924

Discussion

METs is an important clinical variable that has not been previously extracted from medical records through NLP. While ETT is widely used and standardized, its key result, the METs is reported in numerous ways. In the development of an NLP tool for METs extraction, we considered manually creating extraction rules versus machine learning. Our choice of machine learning is informed by our experience with the VA VINCI corpus. Because VHA is a national healthcare system, the VHA text notes are highly varied. On the one hand, it is rich with templates. On the other hand, free text narratives are equally prevalent. To manually create rules for such a diverse corpus, we would need to invest in significant programming effort. With the machine learning approach, we shift a large part of the programming effort to annotation. At the development stage both approaches would likely have cost us similar amounts of time, however the maintenance and enhancement of an annotation dataset is simpler than maintenance of a set of parsing rules in the long run, because the latter requires specialized knowledge of NLP programming. As a result, we decided on the machine learning approach.

Machine learning in NLP is most commonly used for text classification tasks. To extract specific numerical and categorical values, we developed our own REDEx algorithm[14]. The first use case REDEx was applied to was weight value and unit extraction. Comparing to weight, MET is much more complex. First, MET is ambiguous with several meanings (e.g. metastases). Second, MET is reported in many different templates with check boxes, tables, lists, etc. Third, in templates and in free text, the METs value is often not a simple number or even a two-value range. Thus, we modified the REDEx algorithm from the previous study.

Our results show that REDEx was successful in extracting METs values. The precision was 0.9314 and the recall (sensitivity) was 0.8058. The specificity was 0.9562. The F-score and accuracy were 0.8641 and 0.8924 respectively. While we used VHA data for training and testing, the NLP module could easily be customized with additional annotations from additional data sets. This is expected to give straight forward adaptation to non-VHA data, however this needs to be verified in future studies.

A fairly large annotation dataset with 2701 snippets was created in this study. This was necessary as we continue to discover new templates and other new ways METs were described and the recall rates were low at lower sample sizes. Fortunately, METs values or false positives can easily be discerned using short text snippets of 20-30 tokens in length. We were thus able to speed up the annotation by focusing on snippets without reviewing the full-length documents. The total amount of annotation time was about 2 weeks, including development of annotation guidelines.

Our results show a high precision, which is an advantage of this machine learning approach. Machine generated regular expressions do not learn anything wrong because they are continuously validated against the reference set. However, recall can be a challenge because it is difficult to automatically generate regular expressions that can accommodate unseen patterns. Humans are better at anticipating unseen patterns, however it is an onerous task for humans to review the volume of documents that a machine can. We partially addressed this by adding a second tier of regular expressions with a more lenient test function, which improved the recall (which was previously 0.6871) with only minimal harm to the precision (previously 0.9589). We are pursuing further improvement of recall.

A limitation is that a limited set of keywords was used for document retrieval and snippet extraction. Values for METs could potentially occur in contexts other than those within our keyword set. In initial experiments, we included some additional keywords but did not find any examples in which they were associated with METs values. This does not mean METs values in other contexts do not occur, however. Also, we limited snippets to include 20 words before and after the keywords, but is it possible a larger or smaller context would be more accurate. This process assumes that new documents from which values are to be extracted are processed in the same way as the training documents, i.e. split into snippets around keywords prior to application of the regular expressions. Application to entire documents may be desirable, the accuracy of which would need to be evaluated.

In future work, we plan to apply the METs extraction module first to all VHA patients with AF and congestive heart failure (CHF) and then all patients with cardiovascular conditions. The extracted data will become a new structured data variable that can be readily used by researchers as well as clinicians. We also plan to apply REDEx to other cardiovascular measurements. Multiple classification systems are in use that provide more coarse-grained categorization of functional ability[6]. An extension of this study could be to use extracted METs values to derive classifications, or to train REDEx to recognize mentions of those classifications within the text. Extraction of classification values could then be used for estimated METs values. An additional interesting study would be to have a domain expert and regular expression expert manually develop regular expressions based on the same documents and annotation guidelines for comparison of accuracy as well as required effort. Word stemming has been contemplated, however stemming can be a two edged sword. In many clinical circumstances it is a specific disease name or specific label that is of interest, which may occur only in a single word sense in the context of interest. Stemming would increase the false positive rates in these cases. Adding stemming as an option may be an area where future research is warranted.

Acknowledgements

Funding for this project came from the following sources:

HIR 08-374 VA Health Services Research and Development Consortium for Healthcare Informatics Research (CHIR)

HIR 08-204 Veterans Affairs Health Services Research & Development, VA Informatics and Computing Infrastructure (VINCI) aka Center for Scientific Computing

CRE 12-315 Veterans Affairs Health Services Research & Development, CREATE: A VHA NLP Software Ecosystem for Collaborative Development and Integration

References

1. Bierman AS. Functional status: the six vital sign. J Gen Intern Med. 2001;16(11):785-6.
2. Ueshima K, Myers J, Graettinger WF, Atwood JE, Morris CK, Kawaguchi T, et al. Exercise and morphologic comparison of chronic atrial fibrillation and normal sinus rhythm. Am Heart J. 1993;126(1):260-1.
3. Rienstra M, Lubitz SA, Mahida S, Magnani JW, Fontes JD, Sinner MF, et al. Symptoms and functional status of patients with atrial fibrillation: state of the art and future research opportunities. Circulation. 2012;125(23):2933-43.
4. Fleg JL, Pina IL, Balady GJ, Chaitman BR, Fletcher B, Lavie C, et al. Assessment of functional capacity in clinical and research applications: An advisory from the Committee on Exercise, Rehabilitation, and Prevention, Council on Clinical Cardiology, American Heart Association. Circulation. 2000;102(13):1591-7.
5. Jette M, Sidney K, Blumchen G. Metabolic equivalents (METS) in exercise testing, exercise prescription, and evaluation of functional capacity. Clin Cardiol. 1990;13(8):555-65.
6. Fleisher LA, Fleischmann KE, Auerbach AD, Barnason SA, Beckman JA, Bozkurt B, et al. 2014 ACC/AHA Guideline on Perioperative Cardiovascular Evaluation and Management of Patients Undergoing Noncardiac Surgery: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. J Am Coll Cardiol. 2014.
7. Goldman L, Hashimoto B, Cook EF, Loscalzo A. Comparative reproducibility and validity of systems for assessing cardiovascular functional class: advantages of a new specific activity scale. Circulation. 1981;64(6):1227-34.
8. Raphael C, Briscoe C, Davies J, Ian Whinnett Z, Manisty C, Sutton R, et al. Limitations of the New York Heart Association functional classification system and self-reported walking distances in chronic heart failure. Heart. 2007;93(4):476-82.
9. Kokkinos P, Myers J. Exercise and physical activity: clinical outcomes and applications. Circulation. 2010;122(16):1637-48.
10. Bui DD, Zeng-Treitler Q. Learning regular expressions for clinical text classification. J Am Med Inform Assoc. 2014;21(5):850-7.
11. VA Informatics and Computing Infrastructure (VINCI) [html]. United States Department of Veterans Affairs; 2012 [updated Apr. 13, 2012; cited 2012 Sept. 14]. VINCI]. Available from: http://www.hsrd.research.va.gov/for_researchers/vinci.
12. Gundlapalli AV, Redd D, Gibson BS, Carter M, Korhonen C, Nebeker J, et al. Maximizing clinical cohort size using free text queries. Computers in Biology and Medicine. 2015;60:1-7.
13. Visual Tagging Tool: United States National Library of Medicine; 2010 [VTT]. Available from: http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/vtt/current/web/index.html.
14. Murtaugh MA, Gibson BS, Redd D, Zeng-Treitler Q. Regular expression-based learning to extract bodyweight values from clinical notes. Journal of Biomedical Informatics.