Article

Validation of the Total Visual Acuity Extraction Algorithm (TOVA) for Automated Extraction of Visual Acuity Data From Free Text, Unstructured Clinical Records

Douglas M. Baughman[1]*, Grace L. Su[2]*, Irena Tsui[3], Cecilia S. Lee[1]†, and Aaron Y. Lee[1]†

[1] Department of Ophthalmology, University of Washington, Seattle, Washington, USA
[2] Lewis Katz School of Medicine, Temple University, Philadelphia, Pennsylvania, USA
[3] Jules Stein Eye Institute, University of California Los Angeles, Los Angeles, California, USA

Correspondence: Aaron Y. Lee, MD, MSCI, Box 359608, 325 Ninth Avenue, Seattle, WA 98104, USA. e-mail: leeay@uw.edu

Purpose: With increasing volumes of electronic health record data, algorithm-driven extraction may aid manual extraction. Visual acuity often is extracted manually in vision research. The total visual acuity extraction algorithm (TOVA) is presented and validated for automated extraction of visual acuity from free text, unstructured clinical notes.

Methods: Consecutive inpatient ophthalmology notes over an 8-year period from the University of Washington healthcare system in Seattle, WA were used for validation of TOVA. The total visual acuity extraction algorithm applied natural language processing to recognize Snellen visual acuity in free text notes and assign laterality. The best corrected measurement was determined for each eye and converted to logMAR. The algorithm was validated against manual extraction of a subset of notes.

Results: A total of 6266 clinical records were obtained giving 12,452 data points. In a subset of 644 validated notes, comparison of manually extracted data versus TOVA output showed 95% concordance. Interrater reliability testing gave κ statistics of 0.94 (95% confidence interval [CI], 0.89–0.99), 0.96 (95% CI, 0.94–0.98), 0.95 (95% CI, 0.92–0.98), and 0.94 (95% CI, 0.90–0.98) for acuity numerators, denominators, adjustments, and signs, respectively. Pearson correlation coefficient was 0.983. Linear regression showed an R^2 of 0.966 ($P < 0.0001$).

Conclusions: The total visual acuity extraction algorithm is a novel tool for extraction of visual acuity from free text, unstructured clinical notes and provides an open source method of data extraction.

Translational Relevance: Automated visual acuity extraction through natural language processing can be a valuable tool for data extraction from free text ophthalmology notes.

Introduction

With the rise of electronic health records (EHR), an increasing amount of medical data are being stored electronically.[1] This presents the opportunity for large-scale data review in clinical research. However, the quantity of data accessed from EHRs is limited by the time and resources dedicated to manual extraction, traditionally performed by visual inspection of patient charts and manually transcribing data. In addition, many EHR databases, such as the Veteran Affairs National Patient Care Database with more than 20 million eye clinic notes, contain visual acuity (VA) data stored as free text.[2] Therefore, thorough analysis of these large-scale data may be aided by a shift from manual to automated free text extraction.

Automated extraction of EHR data is simplest when dealing with structured data elements.[3] Electronic health data, however, often exists as unstructured free text with inherent ambiguity, loose following of grammatical rules, and lack of easily recognizable data elements.[4] Even in EHR

Table 1. Examples of Free Text VAs and Associated Data Elements

Clinical Record Free Text	Numerator	Denominator	Adjustment	Letters
20/20	20	20		
20/30 −2	20	30	−	2
20/40 +1	20	40	+	1
CF	CF			
HM	HM			
LP	LP			
NLP	NLP			
CF @ 6′	CF	6		
HM @ face	HM	face		
20/30 −	20	30	−	

systems, where click-through and drop-down menus are offered, many physicians opt to use a free text option.[4] Automated conversion of free text into machine-readable data and extraction for research requires the application of computational linguistics and natural language processing. Natural language processing has been applied in numerous disciplines outside biomedical informatics, such as financial market algorithmic trading,[5] social media data mining,[6] sentiment analysis,[7] and machine translation.[7] Within medicine, natural language processing has been investigated in areas such as pathology and radiology where diagnostic data are found within complex prose.[8]

Visual acuity testing is an essential part of an ophthalmologic evaluation.[9] Approximately 14 million Americans over 12 years of age have visual impairment detectable by VA testing.[10] Visual acuity is also an important outcome measure used throughout vision research. Documentation of VA in the clinical record is typically free text with a high degree of variability in the structure of target data elements. Even structured ophthalmology notes usually contain VA data as free text within a structured field, requiring interpretation of the text if the data are to be extracted. Although quantitative data are lacking on the exact nature of VA entry across EHRs, clinical experience shows that few systems require the selection of VA from a predetermined list of options, but instead allow the user to enter free text into a VA field. Using natural language processing, we developed the total VA extraction algorithm (TOVA) to extract best-corrected VA data from free text ophthalmology consultation notes and performed an initial pilot study by validating the results with manually extracted data.

Methods

The study was approved by the University of Washington Institutional Review Board. Research adhered to the tenets of the Declaration of Helsinki and was conducted in accordance with Health Insurance Portability and Accountability Act regulations. We performed a single center, retrospective extraction using structured query language of all electronically available initial ophthalmology consult notes. These were extracted from the underlying database directly from Cerner Powerchart over an 8-year period from July, 2008 to July, 2016 at the University of Washington Medical Center/ Harborview Medical Center in Seattle, WA. A subset of notes had VAs manually extracted for validation of the algorithm. Empty notes with no text or notes that simply referred the reader to a note written in a separate EHR system used by the institution were excluded.

Two study personnel (DB, GS) independently extracted VA data in the traditional fashion (visual inspection and manual copying of data) from a subset of patient notes. Notes were generated by providers typing free text into a text box. These extractors interpreted the free text and converted it to discrete data elements in a spreadsheet comparable to TOVA output (Table 1). A third member of the study team with ophthalmology training (CL) arbitrated discrepancies between the two manually extracted data sets and created a final data set that was used as the gold-standard.

TOVA was created using Ruby (available in the public domain at http://www.ruby-lang.org). A diagram outlining the rule-based natural language processing algorithm, TOVA, created to extract
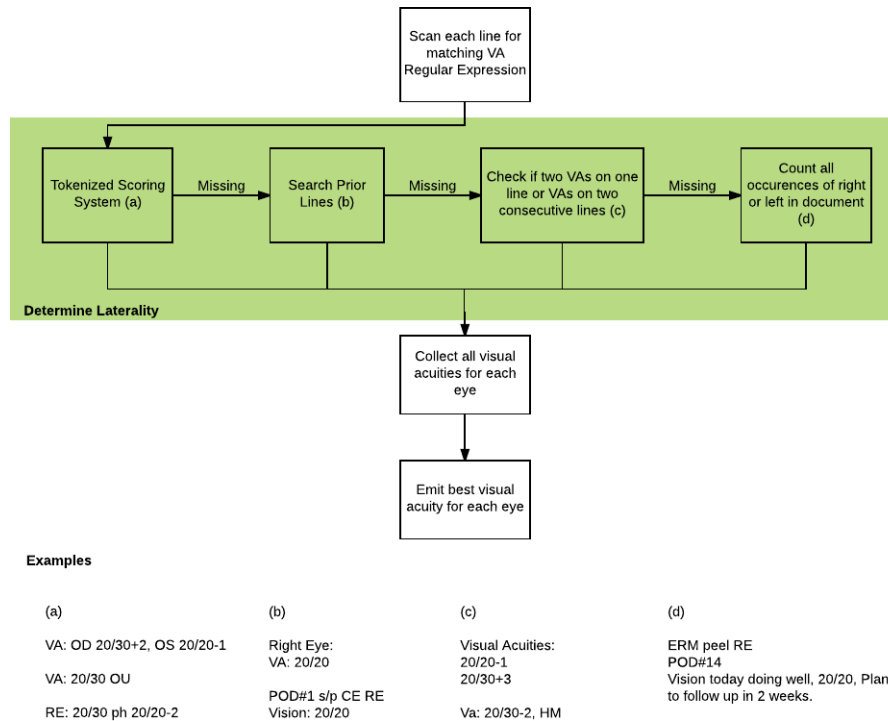
Figure 1. Total VA extraction algorithm logic. A diagram outlining the rule-based natural language processing algorithm, TOVA, for extracting VAs from clinical notes. Stepwise algorithm logic is given above with examples (**a–d**) of free text applications for each step shown below.

VAs from the clinical note is provided in Figure 1. For each line in the clinical note, the following regular expression was applied:

```
/(\s|^|~|:)(20|3E|E)\/\s
*(\d+)\s*([+|-])
*\s*(\d)
*|(HM|CF|LP|NLP)(\W+(@|at|x)
*\s*((\d+)(\s*'|\s*"|\s*in|\s
*ft|\s*feet)*|face)*|$)/
```

A positive match for this regular expression would indicate that a VA was present on the line being evaluated. After a positive match was identified, four strategies were used to evaluate the laterality of the VA found: a tokenized scoring system, searching for laterality in prior lines, determining if two VAs are found either in the same line alone or in two consecutive lines, and counting all the occurrences of right or left in a document. Each step was taken stepwise and if laterality was found then the subsequent steps were not executed.

The tokenized scoring system is diagramed in Figure 2. The line containing the VA was broken into word and punctuation tokens. Each token was scored, with commas and conjunctions receiving a score of 5 and sentence terminators receiving a score of 10.

Synonyms for laterality were determined as follows: word tokens with OD, RE, RIGHT, and R were determined to be about the right eye and word tokens with OS, LE, LEFT, and L were determined to be about the left eye, and OU, BE, BOTH, and BILATERAL tokens were determined to be about both eyes. The scores were determined using the sum of the punctuation tokens between the VA that was identified by the regular expression and the laterality word tokens. The lowest scoring laterality then was assigned to the VA.

If the tokenized scoring system failed, the lines before the line containing the identified VA were searched. Each line was broken into tokens and the first line prior containing a valid word token identifying the laterality was used to assign the VA to an eye.

If searching the prior lines failed to yield a valid laterality, the documentation style may imply the laterality with the right eye VA being listed first and the left eye VA being listed second. The line matching the VA was checked to see if two such matches occurred in the same line, without a prior or subsequent line showing a valid pattern matching the regular expression. The first VA in the line then was assigned to the right eye and the second VA was
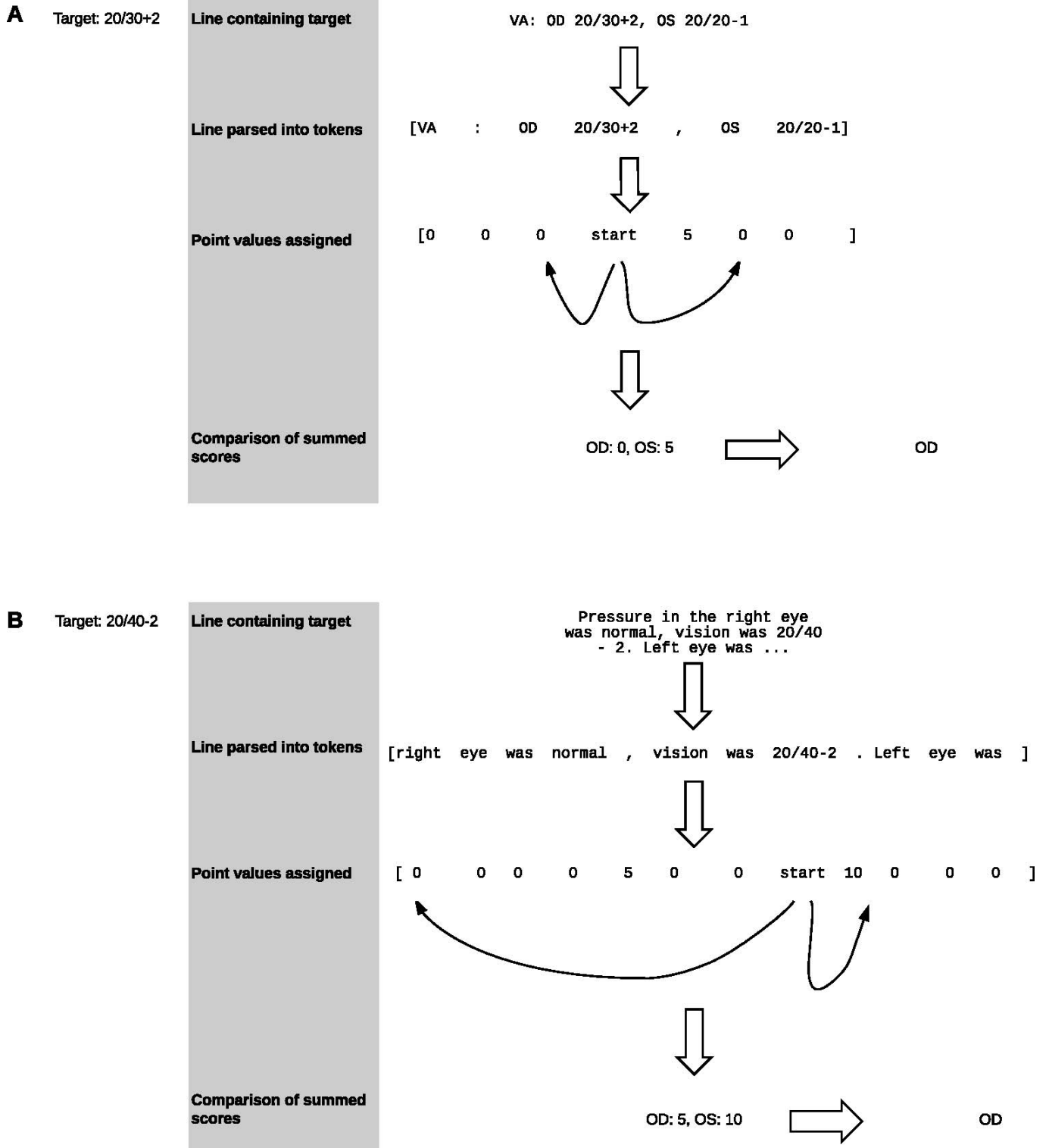
A Target: 20/30+2

Line containing target

VA: OD 20/30+2, OS 20/20-1

Line parsed into tokens

[VA   :   OD   20/30+2   ,   OS   20/20-1]

Point values assigned

[0   0   0   start   5   0   0   ]

Comparison of summed scores

OD: 0, OS: 5   OD

B Target: 20/40-2

Line containing target

Pressure in the right eye
was normal, vision was 20/40
- 2. Left eye was ...

Line parsed into tokens

[right eye was normal , vision was 20/40-2 . Left eye was]

Point values assigned

[0   0 0   0   5   0   0   start 10 0   0   0]

Comparison of summed scores

OD: 5, OS: 10   OD

Figure 2. Tokenized scoring system. Examples of application of the Tokenized scoring system for assigning laterality are given. The line containing VA is parsed into word and punctuation tokens, which are scored, summed, and compared. (A) An example VA target with contralateral laterality token on the same line and separated by a comma. (B) An example VA target with ipsilateral laterality token on the previous line and contralateral laterality token on the following line, separated by a period.

Table 2. Characteristics of Data Elements Extracted from Clinical Records

	N	Percent Exact Match	κ	95% CI
Numerator	1217	98.1%	0.94	0.89–0.99
Denominator	1261	97.9%	0.96	0.94–0.98
Adjustment	127	98.8%	0.95	0.92–0.98
Letters	125	98.7%	0.94	0.90–0.98

N, number of nonblank output values.

assigned to the left eye. If two consecutive lines matched valid patterns and both did not contain valid laterality then the first line VA was assigned to the right eye and the second line VA was assigned to the left eye.

Finally if all the prior methods failed to assign a VA laterality, then the occurrences of all the valid word tokens in the document pertaining to laterality were summed and the highest ranked side was assigned to the VA. Hence, the most frequently mentioned side was assigned as a last resort to determine the laterality.

After all the VAs in the document for each eye were collected, they were converted to logMAR and the best VA was assigned to each eye for the document. Recognition of terms such as "pinhole correction" or "best corrected" was unnecessary given this method of determining best corrected VA. All Snellen VAs were converted to logMAR for analysis. Output VA data was linked to patient identification number, eye, and date of the clinical encounter to aid in downstream clinical research. Output was arbitrarily generated as a tab delimited file that could be imported into a structured query language database, the back end of another EMR, or the Intelligent Research In Sight (IRIS) registry. Visual acuity values count fingers (CF), hand motion (HM), light perception (LP), and no light perception (NLP) were converted to 2.0, 2.4, 2.7, and 3.0, respectively.[11] The exact match rate between manually extracted and algorithm data was calculated for each category. Linear regression of manually extracted versus algorithm data was performed and Pearson's correlation coefficient was calculated. Interrater reliability testing was used to compare manually extracted data to algorithm data with Cohen's κ statistic reported. All analyses were performed using Ruby (available in the public domain at http://www.ruby-lang.org) and R (http://www.r-project.org). The total VA extraction algorithm has been open-sourced under GNU GPLv3 and is now available in the public domain at https://github.com/ayl/vaextractor as a Ruby library.

Results

A total of 12,452 data points was identified from 6266 notes. Mean logMAR VA for the right eye was 0.4507 (median, 0.1761; interquartile ratio [IQR], 0–0.5441) and for the left eye was 0.5078 (median, 0.1761; IQR, 0–0.5441). In the validation subset, 1288 data points were reviewed from 644 notes. Three of the validated notes were excluded due to not having any text or referring to another EMR. In the subset, a total of 644 clinical records from 633 patients yielded 1217 VAs of 1288 data points. In the manually extracted data, VAs ranged from 20/20 to NLP and the most frequent VA was 20/20. All clinical records were written by physicians. Upon arbitration of the two manually extracted data sets, we found 1233 exact matches of 1288 total data elements (95% concordance). The most common reason for discrepancies between the two manual extractors found on arbitration by a third party was VA found in a nonexam section of the note, such as in the assessment and plan.

The total VA extraction algorithm output matched manually extracted data 98.1%, 97.9%, 99.8%, and 98.7% of the time for numerators, denominators, adjustments, and letters, respectively. Kappa statistics were 0.94 (95% confidence interval [CI], 0.89–0.99), 0.96 (95% CI, 0.94–0.98), 0.95 (95% CI, 0.92–0.98), and 0.94 (95% CI, 0.90–0.98) for each data category (Table 2). Pearson correlation coefficient was 0.983. Linear regression showed an R^2 of 0.966 ($P < 0.0001$). A Bland-Altman plot of differences versus averages for paired VAs is shown in Figure 3 and a scatterplot of TOVA-extracted versus manually-extracted data with line of best fit is shown in Figure 4. No systematic discrepancies were found when comparing automated versus manual extraction, as shown in Figure 3.

Discussion

Our study demonstrated that VA data extracted using TOVA correlates to manually extracted data
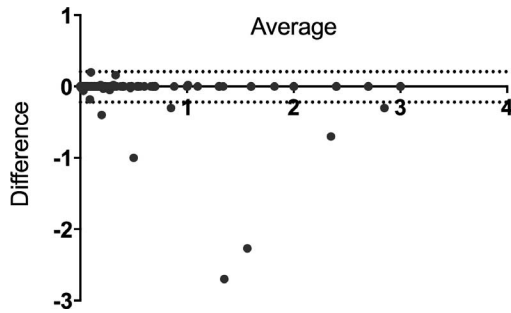
Baughman et al.

Figure 3. Bland-Altman plot. A Bland-Altman plot comparing manually extracted VAs versus TOVA extracted VAs. The average (x-axis) is plotted against the difference (y-axis) of paired VAs. *Dotted lines* represent upper and lower 95% limits of agreement.

Figure 4. Scatterplot. A scatterplot of manually extracted VAs (x-axis) versus TOVA extracted VAs (y-axis). Visual acuities are represented as logMAR values with CF = 2.0, HM = 2.4, LP = 2.7, and NLP = 3.0. Regression line of best fit shown.

with considerable accuracy. Less than one second was required to run TOVA on the corpus of 6266 notes to extract VA and laterality data while the manual extraction of a subset took several days. The total VA extraction algorithm is scalable for much larger datasets, such as the Veteran Affairs National Patient Care Database with more than 20 million free text eye clinic notes.

Our algorithm differs from another recently developed by Mbagwu et al.[12] for extracting VA from EPIC EHR (Epic Systems Corporation, Madison, WI) notes. Their algorithm, written in structured query language, was designed to extract Snellen VAs from structured laterality fields created by the EPIC EHR. It performed keyword searches for text strings within the laterality field that were manually mapped to 1 of 18 defined VA categories (e.g., 20/20, 20/30, and so forth). To assign the best documented VA within a note, they implemented a ranking logic for the 18 categories. They found 5668 unique responses from 298,096 clinical notes, but validated only 100 of these notes using manual chart review and had a match rate of 99%. The total VA extraction algorithm is fundamentally different than the Mbagwu et al.[12] algorithm. Firstly, the use of natural language processing in our algorithm allows for extraction from free text, unlike the Mbagwu et al.[12] algorithm. Their algorithm, while relatively accurate, is designed around structured laterality fields. These fields tell their algorithm which eye the VA belongs to. These fields are not present in many ophthalmology notes and, thus, their algorithm only applies to notes that supply laterality information imbedded in the structure of the note. The total VA extraction algorithm, on the other hand, assigns laterality with the tokenized scoring system which is effective with a block of free
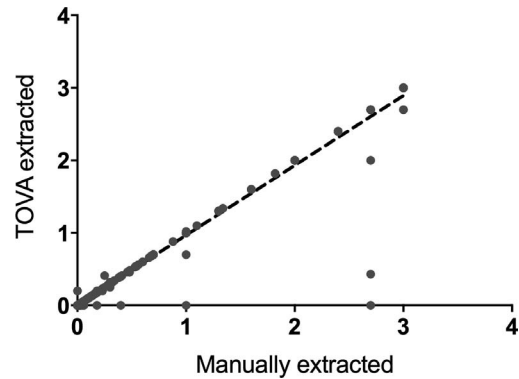
text. Furthermore, since the data within the EPIC EHR laterality fields were free text and their algorithm did not implement natural language processing, they were required to manually map each response to a category, making it difficult to anticipate the full range of possible responses. This also highlights the fact that even in structured notes, VA often is recorded as free text.

In a retrospective study within the Kaiser Permanente Northwest health care system, Smith et al.[13] extracted best corrected VA from 2074 free text notes using a computer program written in Python programming language. They validated their results by manual chart review of 100 notes, but no details about the algorithm logic or results of the validation were reported. Furthermore, their analysis excluded any patient note without VA detected by their algorithm and, therefore, was unable to account for VAs potentially missed.

Natural language processing was used as part of a multimodal approach for extracting cataract cases from broader datasets. In a retrospective review of the Personalized Medicine Research Project (PMRP) cohort, Waudby et al.[14] identified 16,336 cataract patients by combining structured database querying of CPT and ICD-9 codes, natural language processing for data mining of text-based notes, and intelligent character recognition (ICR) of handwritten notes. The results of this combined search were validated by manual extraction of each note. They found a positive predictive value of 95.6% for the combined search when compared to manual extraction. Due to limitations in their automated search, manual extraction was necessary to retrieve data on VA, laterality, and type and severity of cataract. This illustrates the

potential for combining a natural language processing algorithm with other tools for comprehensive automated retrospective review.

Our study has several limitations. We analyzed notes at a single site and, therefore, may not have encountered all variations in VA documentation. However, to the best of our knowledge this is the largest set of notes validated by human extraction and encompasses many styles of note-writers. Multicenter validation of the algorithm is planned in a subsequent study. Our analysis included only inpatient consultation notes, which may be systematically different from outpatient clinic notes. The total VA extraction algorithm is designed to extract from free text notes, and some EHR systems may move toward more structured notes with increased use of drop-down menus or checkboxes. These notes provide more discrete VA data elements and an algorithm designed within that framework may be more accurate. However, EHR systems typically have the capability of exporting notes as free-text, no matter the method of generating the note, and, thus, our algorithm is widely generalizable. While manual extraction currently is the most common method of chart review and was used as the gold standard in our analysis, this method is known to result in transcription error.[3] Indeed, even in our study the interhuman concordance rate was on-par with the concordance of TOVA to final arbitrated data. The total VA extraction algorithm is designed to detect Snellen VAs with imperial measurements and would require modification to detect Snellen metric, logMAR, or other types of VA. Lastly, TOVA was not designed to categorize VAs by the method of measurement (e.g., pinhole aperture testing or unaided VA testing). This is a serious limitation in the current version of the algorithm. The functionality to link the method of measurement to the VA could be added as an extension of the current algorithm. For example, after the best corrected VA is determined, surrounding text then could be searched for the method of measurement and these data could be linked to the VA. Such an extension is planned in an updated version of TOVA.

Despite these limitations, TOVA provides a validated tool for extraction of VA from free text clinical notes, such as those found in large datasets currently available for analysis. The majority of both structured and unstructured notes contain free text VAs making natural language processing a logical approach for extraction. The application of such algorithms has the potential to provide fast, accurate, large-scale data extraction from EHRs allowing more possibilities for future clinical studies.

References

1. Charles D, Gabriel M, Furukawa MF. Adoption of electronic health record systems among US non-federal acute care hospitals: 2008-2012. *ONC Data Brief*. 2013;9:1–9.
2. US Department of Veterans Affairs. 172VA10P2: VHA Corporate Data Warehouse–VA. 79 FR 4377. Updated December 8, 2016. Accessed December 8, 2016.
3. Knake LA, Ahuja M, McDonald EL, et al. Quality of EHR data extractions for studies of preterm birth in a tertiary care center: guidelines for obtaining reliable data. *BMC Pediatr*. 2016; 16:59.
4. Kimia AA, Savova G, Landschaft A, Harper MB. An introduction to natural language processing: how you can get more from those electronic notes you are generating. *Pediatr Emerg Care*. 2015;31:536–541.
5. Zhang W, Skiena S. Trading strategies to exploit blog and news sentiment. In: *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. Menlo Park, CA: AAAI Press; 2010:375–378. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/download/1529/1904/. Accessed September 29, 2016.
6. Li H. Learning to rank for information retrieval and natural language processing. *Synth Lect Hum Lang Technol*. 2014;7:1–121.
7. Balahur A, Turchi M. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Comput Speech Lang*. 2014;28:56–75.

8. Hughes K, Buckley J, Coopey S, et al. The feasibility of using natural language processing to extract clinical information from breast pathology reports. *J Pathol Inform*. 2012;3:23.

9. Feder RS, Olsen TW, Prum BE, et al. Comprehensive adult medical eye evaluation preferred practice pattern guidelines. *Ophthalmology*. 2016; 123:P209–P236.

10. Vitale S, Cotch MF, Sperduto RD. Prevalence of visual impairment in the United States. *JAMA*. 2006;295:2158–2163.

11. Lange C, Feltgen N, Junker B, Schulze-Bonsel K, Bach M. Resolving the clinical acuity categories "hand motion" and "counting fingers" using the Freiburg Visual Acuity Test (FrACT). *Graefes Arch Clin Exp Ophthalmol*. 2009;247:137–142.

12. Mbagwu M, French DD, Gill M, et al. Creation of an accurate algorithm to detect snellen best documented visual acuity from ophthalmology electronic health record notes. *JMIR Med Inform*. 2016;4:e14.

13. Smith DH, Johnson ES, Russell A, et al. Lower visual acuity predicts worse utility values among patients with type 2 diabetes. *Qual Life Res*. 2008; 17:1277–1284.

14. Waudby CJ, Berg RL, Linneman JG, et al. Cataract research using electronic health records. *BMC Ophthalmol*. 2011;11:1.

translational vision science & technology