# eQuality for All: Extending Automated Quality Measurement of Free Text Clinical Narratives

**Steven H Brown MS MD 1,2, Peter L Elkin MD3, S Trent Rosenbloom MD MPH1,2, Elliot Fielstein PhD1, Ted Speroff PhD1,2 1 Department of Veterans Affairs Nashville TN, 2 Vanderbilt University Nashville TN, 3 Mayo Clinic Rochester MN**

***Introduction**: Electronic quality monitoring (eQuality) from clinical narratives may advance current manual quality measurement techniques. We evaluated automated eQuality measurement tools on clinical narratives of veterans' disability examinations.*

***Methods**: We used a general purpose indexing engine to encode clinical concepts with SNOMED CT. We developed computer usable quality assessment rules from established quality indicators and evaluated the automated approach against a gold standard of double independent human expert review. Rules were iteratively improved using a training set of 1446 indexed exam reports and evaluated on a test set of 1454 indexed exam reports.*

***Results**: The eQuality system achieved 86% sensitivity (recall), 62% specificity, and 96% positive predictive value (precision) for automated quality assessment of veterans' disability exams. Summary data for each exam type and detailed data for joint exam quality assessments are presented.*

***Discussion**: The current results generalize our previous results to ten exam types covering over 200 diagnostic codes. eQuality measurement from narrative clinical documents has the potential to improve healthcare quality and safety.*

## Introduction

To date, no reliable comprehensive automated methods for electronically monitoring quality, a process we refer to as eQuality, have been described in the biomedical literature.[1]

The American Health Information Community (AHIC), led by Secretary of Health and Human Services Michael O. Leavitt, is encouraging progress towards electronic quality monitoring by developing a use case and asking the Health Information Technology Standards Panel (HITSP) to create supporting interoperability specifications. The HITSP Population Health Technical Committee is currently accepting comments on its draft quality interoperability specifications and will provide them to the Certification Commission on Health Information Technology (CCHIT) upon finalization. The goal of the AHIC use case, the HITSP interoperability specifications and eQuality in general, is to use health information technology to improve clinical outcomes and safety in the practice of medicine.

A major obstacle to achieving comprehensive eQuality monitoring within the Department of Veterans Affairs is the fact that most electronic health care information in VistA is stored as unstructured free text[2, 3] We believe the situation in healthcare settings outside the VA is analogous. As a result, most current quality monitoring techniques evaluate the subset of health data that is structured or use human reviewers to perform chart abstraction of clinical narratives.[4-6]

Three basic methodologies exist for automatically extracting information from free text that could subsequently be used for computerized quality monitoring and other purposes. These include string matching, computational linguistics (including concept-based indexing), and statistical machine learning techniques.[7-10] We have previously reported on the use of concept-based indexing techniques to aid in automated quality determination of VA spine disability exams.[1]

Veterans' disability exams make up an important subset of health data for eQuality determination. In fiscal year 2006, VA conducted over 800,000 disability exams and distributed over $34 billion in disability benefits to approximately 2.9 million veterans, including an ever increasing number of young veterans returning from Iraq and Afghanistan. VA created the Compensation and Pension Exam Program (CPEP) in 2001 in order to address disability exam quality on a national scale in recognition of the fact that high quality disability exams are important for accurate disability entitlement decisions.

In the current study, we extended our eQuality monitoring evaluation to cover the ten most commonly requested veterans' disability exam types. The 'top ten' exam types cover 65% of VHA C&P exam workload.

## Methods

*Data Sets*

VA Compensation and Pension (C&P) exams are conducted according to one of 59 protocols (i.e., heart, skin, diabetes, mental disorders). Each month CPEP subjects approximately 140 of each of the ten most commonly requested exam protocol types to double independent human expert review. Reviews are performed to measure examiners' compliance with established quality measures.[4] To facilitate this quality measurement process, CPEP staff downloads all VHA compensation and pension exams released electronically during the prior month by examining sites nationwide. Between June and December 2007, VHA electronically released an average of nearly 78,000 exams per month from the 128 VistA systems in production at medical centers nationwide.

We extracted all of the 'top ten' exam text-based reports that were released by VHA examiners during the months of December 2005 and January 2006 and reviewed by trained CPEP reviewers the following months. We also extracted the expert consensus quality indicator review results for each exam to serve as a gold standard for algorithmic eQuality classification. We arbitrarily assigned exams released in December 2005 (n = 1,446) to be the study training set and exams released in January 2006 (n = 1,454) to be the study test set. Our methods for assessing quality from free text involve three major steps: document indexing, rule formulation and rule evaluation.

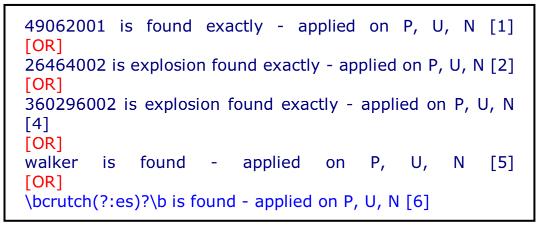
*Document Indexing*

We applied concept-based indexing techniques to each of the 2900 VHA disability exams in the study. The NLP indexing engine and review system we used (LingoEngine and GoReview, LingoLogix, Dallas Texas) are commercially available products associated with the research version we previously described.[11] The research version had a sensitivity of 99.7% and a specificity of 97.9% for SNOMED CT encoding of Mayo problem statements. We used the indexing engine to process words and phrases found in narrative exam reports into SNOMED CT (Jan 2003) encoded concepts through four steps: 1) parsing the documents into sections and sentences 2) normalizing words; 3) mapping words to concepts; 4) mapping single concepts to more complex concepts.

*Quality Rule Formulation*

The foundation of the automated quality assessment rules evaluated in this study are the set of established VA C&P exam quality indicators used by trained expert C&P exam quality reviewers around the country. The quality indicators were developed by expert clinical and disability rating panels and have been used nationwide since 2003 in support of a VHA-wide performance measure. The rules include two types of quality indicators – exam specific quality indicators and core quality indicators. Exam specific quality indicators apply to one exam type (e.g., documentation of noise exposure in the military applies only to audiology exams) and core quality indicators that apply to all exam types (e.g., "did the exam address all issues requested?"). Joint exam specific indicators are presented in  table 3.

For the current study we developed and evaluated a computer usable rule for each exam specific quality indicator associated with the ten most commonly requested exam types. We translated each quality indicator into a computer usable rule via a series of steps. The first step was to map clinical concepts found in the quality indicator to terminological concepts in SNOMED CT (June 2003).[12] Mappings to single CT concepts or to CT concepts and all their descendants ("concept explosions") were permitted. When appropriate concepts could not be found in CT we used simple term matching via regular expressions. We composed draft computer usable rules by grouping mapped CT concepts, concept explosions and terms with Boolean operators (Figure 1). Rules and document indices are created using the same language and logic.



```
49062001 is found exactly - applied on P, U, N [1]
[OR]
26464002 is explosion found exactly - applied on P, U, N [2]
[OR]
360296002 is explosion found exactly - applied on P, U, N
[4]
[OR]
walker is found - applied on P, U, N [5]
[OR]
\bcrutch(?:es)?\b is found - applied on P, U, N [6]
```

**Figure 1.** A portion of a rule regarding the use of assistive devices for walking.

We next applied the draft rules for each exam type to exam reports that had previously undergone gold standard quality review and indexing. We compared rule outputs to gold standard quality assessments. When human and automated quality indicator assessments differed we reviewed the exam narrative and adjusted the computer executable rule accordingly. We repeated the cycle of rule application, results review and rule modification on the training sets until further gains in sensitivity and specificity could no longer be achieved. The final rule set for each exam type was then applied to the test set of human reviewed and indexed exams. We report summary counts of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) classifications and the resulting sensitivity (sen), specificity (sp), positive likelihood ratio (plr), positive predictive value (ppv) and negative predictive value (npv) for each of the ten exam types. We report the same statistics for each of the individual joint exam specific quality assessment rules as an example of the larger data set. .

## Results

Overall, we composed 95 rules using 396,175 concepts, Y and 2,203 strings (Table 1). On the training set overall sensitivity (recall) was 87%, specificity was 61% and positive predictive value (precision) was 96%. On the test set overall sensitivity (recall) was 86%, specificity was 62% and positive predictive value (precision) was 96% (Table 2). Table 3 details system performance for each of the 10 joint exam specific quality indicators. Column one includes the text of the quality indicator used by expert reviewers to determine gold standard quality for this study and for administration of a national VA disability exam quality performance measure.

| Exam | Doc Count | Rule | conc | conc exp | str ptrn | TP | TN | FP | FN | sen | sp | p-lr | ppv | npv |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Audio | 147 | 7 | 37 | 13,810 | 67 | 973 | 5 | 3 | 48 | 95% | 63% | 254% | 100% | 9% |
| Eye | 143 | 7 | 43 | 41,977 | 67 | 856 | 31 | 20 | 94 | 90% | 61% | 230% | 98% | 25% |
| Feet | 143 | 10 | 83 | 41,296 | 244 | 1,013 | 90 | 47 | 110 | 90% | 66% | 263% | 96% | 45% |
| GenM | 145 | 10 | 145 | 118,047 | 296 | 1,110 | 115 | 81 | 245 | 82% | 59% | 198% | 93% | 32% |
| iPTSD | 145 | 10 | 118 | 14,943 | 199 | 1,102 | 61 | 36 | 251 | 81% | 63% | 219% | 97% | 20% |
| Joints | 142 | 10 | 92 | 38,338 | 250 | 1,026 | 90 | 45 | 119 | 90% | 67% | 269% | 96% | 43% |
| Mental | 144 | 10 | 119 | 18,456 | 335 | 1,090 | 66 | 42 | 232 | 82% | 61% | 212% | 96% | 22% |
| rPTSD | 147 | 10 | 136 | 46,169 | 374 | 1,147 | 66 | 49 | 208 | 85% | 57% | 199% | 96% | 24% |
| Skin | 144 | 10 | 55 | 39,458 | 157 | 860 | 47 | 40 | 113 | 88% | 54% | 192% | 96% | 29% |
| Spine | 146 | 11 | 74 | 23,681 | 214 | 989 | 63 | 41 | 106 | 90% | 61% | 229% | 96% | 37% |
| Total | 1446 | 95 | 902 | 396,175 | 2203 | 10,166 | 634 | 404 | 1526 | 87% | 61% | 223% | 96% | 29% |

**Table 1**. Rule characteristics and training set performance for VA Disability exams. The "i" and "r" refer to initial and review post traumatic stress disorder exams. The number of rules, concepts (conc), exploded concepts (conc exp) and strings or patterns (str ptrn) for each exam are shown in columns 3-6.

| Exam | Docs Count | Rules | TP | TN | FP | FN | sen | sp | p-lr | ppv | Npv |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Audio | 146 | 7 | 989 | 63 | 41 | 106 | 90% | 61% | 229% | 96% | 37% |
| Eye | 144 | 7 | 820 | 40 | 22 | 112 | 88% | 65% | 248% | 97% | 26% |
| Feet | 145 | 10 | 943 | 58 | 35 | 104 | 90% | 62% | 239% | 96% | 36% |
| GenM | 145 | 10 | 1096 | 104 | 70 | 303 | 78% | 60% | 195% | 94% | 26% |
| iPTSD | 146 | 10 | 1150 | 41 | 24 | 245 | 82% | 63% | 223% | 98% | 14% |
| Joints | 144 | 10 | 1086 | 84 | 40 | 100 | 92% | 68% | 284% | 96% | 46% |
| Mental | 146 | 10 | 1093 | 61 | 35 | 251 | 81% | 64% | 223% | 97% | 20% |
| rPTSD | 147 | 10 | 1131 | 92 | 60 | 187 | 86% | 61% | 217% | 95% | 33% |
| Skin | 144 | 10 | 917 | 58 | 39 | 126 | 88% | 60% | 219% | 96% | 32% |
| Spine | 147 | 11 | 1068 | 66 | 42 | 89 | 92% | 61% | 237% | 96% | 43% |
| **Total** | 1454 | 95 | 10293 | 667 | 408 | 1623 | 86% | 62% | 228% | 96% | 29% |

**Table 2.** Test set performance for each of the ten most commonly requested VA Disability exams.

| **Joint Exam Quality Indicators** | **TP** | **TN** | **FP** | **FN** | **sen** | **sp** | **p-lr** | **ppv** | **npv** |
|---|---|---|---|---|---|---|---|---|---|
| 1. Does report note subjective complaints? | 131 | 0 | 0 | 0 | 100% | -N/A- | -N/A- | 100% | -N/A- |
| 2. Does report describe need for assistive devices? | 123 | 3 | 1 | 4 | 97% | 75% | 387% | 99% | 43% |
| **3.** Does the report describe the effects of the condition on the veteran's usual occupation? | 75 | 25 | 4 | 27 | 74% | 86% | 533% | 95% | 48% |
| 4. Does report describe effects of the condition on the veteran's routine daily activities? | 109 | 10 | 3 | 9 | 92% | 77% | 400% | 97% | 53% |
| **5.** Does report provide the active range of motion in degrees? | 127 | 1 | 3 | 0 | 100% | 25% | 133% | 98% | 100% |
| 6. Does the report state whether the joint is painful on motion | 85 | 17 | 10 | 19 | 82% | 63% | 221% | 89% | 47% |
| 7. Does the report address additional limitation following repetitive use? | 99 | 15 | 7 | 10 | 91% | 68% | 285% | 93% | 60% |
| 8. Does the report describe flare-ups? | 103 | 11 | 6 | 11 | 90% | 65% | 256% | 94% | 50% |
| 9. Does report address instability of knee? | 109 | 2 | 2 | 18 | 86% | 50% | 172% | 98% | 10% |
| 10. Does the report include results of all conducted diagnostic and clinical tests? | 125 | 0 | 4 | 2 | 98% | 0% | 98% | 97% | 0% |
| **totals** | **1086** | **84** | **40** | **100** | **92%** | **68%** | **284%** | **96%** | **46%** |

**Table 3**. eQuality performance for test set joint disability exam specific quality indicators

## Discussion

Overall, the eQuality tool achieved 86% sensitivity (recall) 62% specificity and 96% positive predictive value (precision) for automated quality assessment from test set narratives representing the ten most commonly requested veterans' disability exams, a generalization of our previous results evaluating spine exams.

We believe this is an important step towards demonstrating the general utility of using coded concepts extracted from unstructured text for automated quality measurement for four reasons. First, system performance remained stable between training and test sets. Second, the ten most common exams cover a wide variety of physical and mental disorders and body systems. Excluding the general medical evaluation, an umbrella screening exam, the studied exam types cover over 200 diagnostic conditions. Third, most of the exam specific quality indicators asked were of similar or greater complexity than the questions posed in the I2B2 smoking status challenge featured in the January-February 2008 issue of JAMIA.[13-18] Finally, our approach relies on the same general purpose indexing tool and ontology for all exam types studied. We did not create special purpose indexing machines and ontologies for each exam type. New rules for other quality measures could be developed and executed without re-indexing the documents.

It is interesting to note that system performance appears to be better for exams of physical conditions than for exams of psychiatric conditions. It is not clear whether this performance difference is a function of the quality indicators, SNOMED CT, the indexing engine, or our ability to create computer executable rules. This is an important area that merits further study.

The current study evaluated ten veterans' disability exam types but may not generalize to other types of clinical documentation. In previous work, we found that 96.2% of elements on the general medical disability evaluation were not specific to disability evaluation. That said, veterans' disability exams contain extensive historical (including treatment), exam and assessment data but do not place a heavy focus on treatment planning.

Automated eQuality monitoring requires less than a minute per record and promises to be less expensive than abstraction by human experts. The availability of a fast, accurate and inexpensive mechanism for quality measurement could greatly expand our ability to guide quality improvement with timely data. Additional studies are planned to evaluate implementations of the tool.

It is our hope that eQuality monitoring can be extended beyond disability exams to a wider array of healthcare applications. eQuality solutions based on a core data infrastructure of encoded data extracted from free-text and validated automated quality assessment rules could equip healthcare organizations to monitor their care quality in near real time. Subsequently, it may be possible to provide clinical reminders and guidance to clinicians at the point and time of care based on analysis of past and newly typed records. Although much work remains to be done to reach these objectives, we are optimistic that recent advances in electronic health record deployment, formal terminologies, and text processing technologies can speed progress in this important area.

Acknowledgements: This work was supported in part by grants CDCPH000022, HK00014,VA SAF03-223-3, and NLM: K22 5LM008576-03. The authors would also like to acknowledge the contributions of Mr. Constantin Mitocaru (Technology Developer at LingoLogix), Ms. Glenda Taylor and the VA CPEP reviewers.

## References

1. Brown SH, Speroff T, Fielstein EM, Bauer BA, Wahner-Roedler DL, Greevy R, et al. eQuality: electronic quality assessment from narrative clinical reports. Mayo Clin Proc. 2006 Nov;81(11):1472-81.
2. Brown SH, Lincoln MJ, Groen P, Kolodner RM. VistA: The U.S. Department of Veterans Affairs National Scale Hospital Information System. International Journal of Medical Informatics. 2003;69(2-3):135 - 56
3. Kolodner RM, editor. Computerizing large integrated health networks: the VA success. New York: Springer-Verlag; 1997.
4. Weeks WB, Mills PD, Waldron J, Brown SH, Speroff T, Coulson LR. A model for improving the quality and timeliness of compensation and pension examinations in VA facilities. J Healthc Manag. 2003 Jul-Aug;48(4):252-61; discussion 62.
5. Berry K. Legislative forum: HEDIS 2.0: a standardized method to evaluate health plans. J Healthc Qual. 1993 Nov-Dec;15(6):42
6. Campbell S. Outcomes-based accreditation evolves slowly with JCAHO's Oryx initiative. Health Care Strateg Manage. 1997 Apr;15(4):12-3.
7. Baud RH, Lovis C, Ruch P, Rassinoux AM. Conceptual search in electronic patient record. Medinfo. 2001;10(Pt 1):156-60.
8. Baud RH, Rassinoux AM, Ruch P, Lovis C, Scherrer JR. The power and limits of a rule-based morpho-semantic parser. Proc AMIA Symp. 1999:22-6.
9. Goldman JA, Chu WW, Parker DS, Goldman RM. Term domain distribution analysis: a data mining tool for text databases. Methods Inf Med. 1999 Jun;38(2):96-101.
10. Cooper GF, Miller RA. An experiment comparing lexical and statistical methods for extracting MeSH terms from clinical free text. J Am Med Inform Assoc. 1998;5(1):62-75.
11. Elkin PL, Brown SH, Husser CS, Bauer BA, Wahner-Roedler D, Rosenbloom ST, et al. Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. Mayo Clin Proc. 2006 Jun;81(6):741-8.
12. Stearns MQ, Price C, Spackman KA, Wang AY. SNOMED clinical terms: overview of the development process and project status. Proc AMIA Symp. 2001:662-6.
13. Clark C, Good K, Jezierny L, Macpherson M, Wilson B, Chajewska U. Identifying smokers with a medical extraction system. J Am Med Inform Assoc. 2008 Jan-Feb;15(1):36-9.
14. Cohen AM. Five-way smoking status classification using text hot-spot identification and error-correcting output codes. J Am Med Inform Assoc. 2008 Jan-Feb;15(1):32-5.
15. Heinze DT, Morsch ML, Potter BC, Sheffer RE, Jr. Medical i2b2 NLP smoking challenge: the A-Life system architecture and methodology. J Am Med Inform Assoc. 2008 Jan-Feb;15(1):40-3.
16. Savova GK, Ogren PV, Duffy PH, Buntrock JD, Chute CG. Mayo clinic NLP system for patient smoking status identification. J Am Med Infrm Assoc. 2008 Jan-Feb;15(1):25-8.
17. Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc. 2008 Jan-Feb;15(1):14-24.
18. Wicentowski R, Sydes MR. Using implicit information to identify smoking status in smoke-blind medical discharge summaries. J Am Med Inform Assoc. 2008 Jan-Feb;15(1):29-31.