Automatic Quality of Life Prediction Using Electronic Medical Records

Serguei Pakhomov, PhD[1], Nilay Shah, PhD[2], Penny Hanson[2], Saranya Balasubramaniam[2], and Steven A. Smith, MD[3]

[1] *Pharmaceutical Care and Health Systems, University of Minnesota, MN, USA*
[2] *Health Care Policy & Research, Mayo College of Medicine, Rochester, MN, USA*
[3] *Endocrinology, Mayo College of Medicine, Rochester, MN, USA*

Abstract

Health related quality of life (HRQOL) is an important variable used as a risk factor for prognosis and as an outcome in clinical studies and for quality improvement. We explore the use of a general purpose natural language processing system (Metamap) in combination with Support Vector Machines (SVM) for predicting patient responses on standardized HRQOL assessment instruments from the text of physician's notes. We surveyed 669 patients in the Mayo Clinic diabetes registry using two instruments designed to assess functioning: EuroQoL5D and SF36/SD6. Clinical notes for these patients were represented as sets of medical concepts using Metamap. SVM classifiers were trained using various feature selection strategies. The best concordance between the HRQOL instruments and automatic classification was achieved along the "pain" dimension (positive agreement – .76, negative agreement – .78, kappa – .54) using Metamap. We conclude that clinician's notes may be used to develop a surrogate measure of patient's HRQOL status.

Introduction

Patient's health related quality of life (HRQOL) is an important variable used both as a risk factor for prognosis and as an outcome in clinical studies and quality of care initiatives. Typically, HRQOL information is collected directly from patients using standardized survey instruments including questionnaires and visual analog scales. Administering these instruments is a time consuming process not practical in an already time constrained clinical encounter. Furthermore, their clinical use is subject to responsiveness bias and is not easily linked to the patient encounter permitting efficient measurement of patient-reported outcomes longitudinally. In addition, it is not practical to collect HRQOL information for large populations. Physician reports contained in the Electronic Medical Records (EMR) may serve as an alternative source of HRQOL information but require natural language processing (NLP) technology to extract this information and currently have not been validated for this purpose.

Materials and Methods

Participants : We have previously reported the results of the UNITED Planned Care Trial; clinicaltrials.gov: NCT00421850, a population based randomized controlled trial assessing the value of virtual consultations in the care of patients seen in primary care for diabetes.[1] The Mayo Foundation Institutional Review Board approved the study procedures and all patients participating gave written informed consent and research authorization.

The patient population for this trial was representative of the six primary care family and internal medicine practices affiliated with Mayo Clinic, a large academic medical center in Rochester, Olmsted County, Minnesota, USA. Mayo Clinic provides primary care to local residents, including over 5000 patients with diabetes (approximately 6% of the population) whose characteristics are similar to those of US non-Hispanic whites.[2]

HRQOL Assessment Instruments: To assess the functional health status of participating individuals we mailed the EuroQoL5D (EQ5D) and the Medical Outcomes Study Short Form (SF36) to 669 patients upon first referral for randomization to receive virtual consultations (recruitment dates July 2001- December 2003). A second mailing followed 3 weeks later to non-responders with an overall response rate of 67% (n=447).

The EQ5D measures functional health status on three levels along 5 dimensions: pain, mobility, usual activities, self-care and depression/anxiety. The SF36 consists of 36 questions and 8 domains. Both instruments have been extensively validated in diverse population groups with chronic disease to include diabetes.[3, 4] In addition, Brazier et al have validated an SF6D index collapsing the 36 questions from the SF36 to similar dimensions as the EQ5D.[5] The six dimensions for the SF6D are physical functioning (6 levels) corresponding to EQ5D mobility, role limitations (4 levels) and social functioning (5 levels) corresponding to EQ5D usual activities, pain (6 levels) corresponding to EQ5D pain/discomfort, mental health (5 levels) corresponding to EQ5D depression/anxiety, and vitality (5 levels) without an EQ5D counterpart. Both the EQ5D and the SF6D are two of the most widely used measures for choice-based methods for valuation of health states.[6-8]

The Likert-style scales of the EQ5D and SF6D indexes were dichotomized into "normal" and "abnormal" categories for the purposes of this study. We experimentally determined cut-points to distinguish between responses indicating "normal" and "abnormal" functioning by plotting the distribution of the multi-valued responses along each dimension and then manually finding a point that bisects the data into roughly equal proportions.

Clinical Notes and Mayo Clinic EMR: Mayo Clinic providers have been documenting each patient encounter electronically since 1994. Currently clinical notes are either self entered or dictated and stored in electronic format (Mayo Clinic EMR) comprising a dataset of over 25 million in-patient and out-patient notes. These notes are in compliance with the American National Standards Institute Clinical Document Architecture, a for clinical documentation.[9] All clinical notes for the 447 patients dictated between July 2001 and September 2004 were used in this study.

Machine Learning: We used a Support Vector Machine (SVM) classifier to predict patients' responses along the EQ5D and SF6D dimensions. An SVM is set of automatic classification algorithms used in medical text categorization.[10-14] SVMs classify text by "learning" the optimal shape of a multi-dimensional partition that separates the data points into two categories.[15] The partition is iteratively fitted to an abstract multi-dimensional space where each dimension is represented by a predictive feature. Thus an SVM in principle is similar to a neural network[15-17] where the goal is to find the most optimal analytical solution that maps an input pattern of predictive covariates to the correct category in the output. For this study, we used a WEKA SVM implementation of the sequential minimal optimization algorithm (SMO) classifier[18] with the default parameter settings – attribute normalization, polynomial kernel with exponent of 1.0, and complexity of 1.0. No parameter optimization was performed for this study.

Feature extraction: As with most machine learning approaches, SVMs require training data where each training sample consists of a set of predictive features and is labeled with the correct category the SVM is expected to "learn." We experimented with two types of predictive feature representation. The first type relied on a simple "bag-of-words" method where the text of clinical notes was parsed into single words and each unique word occurring more than three times in all samples was used to represent the vocabulary of predictive features. In addition to the frequency cutoff of 3, we used a list of 134 "stopwords." The stopwords included function words (e.g., a, the, on, in, is, that, etc.) frequently removed from the predictive feature space in text categorization problems. The second type relied on representing each clinical note as an unordered set of medical concepts (a "bag-of-concepts" approach). To identify medical concepts we used the publicly available implementation of Metamap (MMTx-2.4b). Metamap is a general purpose NLP system developed at the National Library of Medicine for biomedical applications[19, 20]. It operates by identifying noun, verb and prepositional phrases with the help of a minimal commitment parser[21] and bringing lexical and morphological variants of medical terms to a standard form. It also uses linguistic principles to map the different types of phrases to the Unified Medical Language System (UMLS) Metathesaurus. The latter represents a system of vocabularies, nomenclatures and ontologies (e.g., multiple revisions of the International Classification of Diseases (ICD), Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT), and RxNorm among over 100 other sources). For this study we maintained the default settings on Metamap and set the confidence score to 900, thus the text of clinical notes would only be mapped to concepts if Metamap was "confident" about the accuracy of the mapping.

Feature selection: The feature extraction methods described in the previous section result in very large numbers (> 10,000) of predictive features. Reducing the feature space to non-redundant features results in improved classification accuracy and helps avoid overfitting of the classifiers.[22] In this study, we experimented with a hybrid method for selecting predictive features that combines two existing methods: Correlation-based Feature Subset (CFS) and Information Gain (IG).

The CFS method uses a heuristic to evaluate the merit of feature subsets with respect to classification categories and the correlation between the features. While this method tends to produce accurate classifiers, it does require considerable time for computation with very large feature spaces. Information gain of a predictive feature can be defined as the difference between the amount of information necessary to specify the outcome without any predictor variables and the amount of information contributed by the feature.[15] The greater the information gain of a predictor, the more discriminative power the predictor will have in the classifier. The advantage of the IG feature selection method is its efficiency necessary when working with large feature spaces. We experimented with the IG selection method by itself and in a cascade with the CFS method where we first select 1000 features with the highest information gain values and then further reduce this subset to a much smaller (< 200) subset using the CFS approach.

Automatic Classification: We experimented with two feature extraction and two feature selection methods in addition to the baseline with all features present. Thus, we trained and tested a total of six SVM classifiers for each dimension of the two HRQOL instruments resulting in a total of 66 classifiers.

Manual Classification: In addition to the automatic classifiers, we categorized medical records of 169 randomly selected patients from the group of 447 using two human experts with experience in diabetes care – an endocrinologist and a Diabetes Electronic Management System liaison (S.S. and P.H.). The reviewers manually examined patient records and classified them along the five EQ5D dimensions. This classification was performed to isolate the effects of the algorithms from the effects of the information contained in the text of clinical reports as detailed further in the Results and Discussion section.

Statistical Methods: Our evaluation is based on the comparison between the classifications produced by SVMs and the patient responses to the questions in the HRQOL assessment instruments. Since neither one can be considered a reference standard, we report the results in terms of agreement rather than sensitivity and specificity. Kappa statistic is a standard for measuring agreement; however, it has been shown to be sensitive to imbalances in the marginal totals of comparisons involving two categories,[23] which is the case in this study. Positive and negative agreement measures have been proposed as a way to ensure the correct interpretation of kappa values[24] and have been used previously to assess the agreement between patient reported information and the medical record.[25] The positive agreement (Ppos) is a ratio of the concordances in positive responses (TP) to the difference between the concordances in positive (TP) and negative (TN) responses added to the total number of samples according to the formula in (1):

$$(1) \qquad Ppos = \frac{2TP}{N + (TP - TN)}$$

The negative agreement (Pneg) is a ratio of the concordances in negative responses (TN) to the difference between the concordances in positive (TP) and negative responses (TN) subtracted from the total number of samples according to the following formula in (2):

$$(2) \qquad Pneg = \frac{2TN}{N - (TP - TN)}$$

The negative and positive agreement were calculated based on the averaged results obtained with 10-fold crossvalidation, a standard technique for evaluating automatic classifiers.[15]

The agreement between multi-level patient responses on EQ5D and manual record review was computed using the intra-class correlation coefficient (ICC). Following Shrout and Fleiss[26] a two-way ANOVA method based on average measures was used to compute the ICC.

Results and Discussion

Cut-points for HRQOL Instruments: We dichotomized the scales in this manner to reduce the number of categories for machine learning in this preliminary investigation. The distribution of the Likert-scale values for SF6D and EQ5D indexes is shown in Table 1 and Table 2. The values in boldface font indicate the cut-points where, for example, for the *Pain* dimension in Table 1, we will consider scale values of 1 and 2 as "normal" while the rest – "abnormal." The *Self-care* dimension in Table 2 has the value of 1 for 92% of the patients, indicating that there are not enough "abnormal" samples for this dimension to be used in the evaluation of the SVM algorithms. Thus we excluded this dimension from further analysis.

Table 1 Distribution of patient responses on SF6D composite index along six HRQOL dimensions

	SF6D dimension (N=447)					
Scale	Phys	Role*	Soc^	Pain	Ment^	Vital^
1	.15	**.49**	**.59**	.17	.30	.03
2	**.39**	.19	.20	**.25**	**.32**	**.34**
3	.24	.12	.14	.26	.30	.36
4	.09	.20	.05	.19	.06	.18
5	.10	N/A	.01	.12	.02	.10
6	.03	N/A	N/A	.00	N/A	N/A

*Dimension has 4 levels, ^Dimension has 5 levels

Feature Extraction: We retrieved 24,744 clinical notes for the 447 study participants and converted them to bag-of-words and bag-of-concepts feature vectors. The bag-of-words feature space consisted of 27,403 words, while the bag-of-concepts space consisted of 10,735 concepts.

Table 2 Distribution of patient responses on EQ5D composite index along five HRQOL dimensions

	EQ5D dimension (N=447)				
Scale	Mobil	Self	Usual	Pain	Depres
1	**.58**	**.92**	**.63**	**.29**	**.63**
2	.41	.08	.35	.68	.36
3	.00	.00	.01	.04	.02

Feature Selection: The IG feature selection method was set to produce a fixed number of features – top 1000. Subsequent application of the CFS method resulted in further reduction. The sizes of feature subsets are displayed in the shaded areas of Table 3 and Table 5.

Agreement for automatic classification: The agreement statistics for each of the classifiers in this study are shown in Table 3 and Table 5. On average, across all 6 dimensions of SF6D, the bag-of-concepts method with CFS feature selection has the best agreement (k = .52) with patients' responses of the other methods. The best agreement is achieved along the *Pain* (k = .54) and the *Mental* (k = .60) dimensions. Furthermore, the positive and negative agreement along these two dimensions is more balanced than, for example, along the *Role* dimension where Kappa is relatively high (k = .50) but the positive agreement is .15 higher than the negative agreement. This imbalance indicates that the classifier was concordant on "normal" HRQOL samples but discordant on many of the "abnormal" samples.

The results of agreement with EQ5D are similar to those for SF6D with respect to the distinction between the bag-of-words and the bag-of-concepts approach. The results are not as clear with respect to the feature selection methods. The results in Table 5 indicate that agreement is generally better with the bag-of-concepts approach (k = .52 vs. k = .42). However, while the CFS feature selection method showed a substantial improvement over the baseline (k = .48 vs. k = .23), it performed worse than the IG method by .04. This is due to the fact that the classifier trained using the CFS method had a very high level of agreement on the "abnormal" samples (Pneg = .87) but a much lower agreement on "normal" samples (Ppos = .46). The best agreement is achieved with the IG method along the *Pain* dimension (k = .63, Ppos = .74, Pneg = .88). The worst agreement is found along the *Depression* dimension (k = .47, Ppos = 84, Pneg = .82).

Table 3 Agreement between HRQOL dimensions of the SF6D index and SVM classification of clinical notes

	Feature extraction Method					
	Bag-of-words			Bag-of-concepts		
	Feature selection Method					
Dimension	NFS	IG	CFS	NFS	IG	CFS
Pain			87*			134
Kappa	.28	.38	.42	.20	.44	.54
Ppos	.60	.67	.69	.55	.70	.76
Pneg	.68	.71	.72	.65	.73	.78
Physic.			96			127
Kappa	.33	.31	.47	.34	.45	.49
Ppos	.73	.72	.80	.72	.77	.80
Pneg	.58	.58	.66	.62	.68	.68
Role.			85			118
Kappa	.22	.33	.43	.16	.40	.50
Ppos	.79	.80	.85	.76	.83	.87
Pneg	.42	.52	.56	.39	.57	.62
Viality			114			124
Kappa	.11	.39	.45	.13	.58	.47
Ppos	.39	.63	.67	.46	.75	.63
Pneg	.71	.75	.77	.67	.83	.84
Social			89			119
Kappa	.19	.36	.40	.29	.39	.50
Ppos	.71	.77	.80	.75	.77	.83
Pneg	.46	.58	.59	.53	.61	.66
Mental			107			162
Kappa	.12	.53	.50	.09	.52	.60
Ppos	.74	.85	.85	.71	.85	.88
Pneg	.36	.68	.65	.37	.66	.71
Mean Kappa	.21	.38	.45	.20	.46	.52

* Number of predictive features selected by the CFS algorithm

Agreement with manual classification: The agreement results are presented in Table 4. The inter-rater agreement between SS and PH is in the "good" category[26] for *Pain*, *Usual Activities* and *Depression* dimensions; however, the agreement is only "moderate" on *Mobility*. The comparison with EQ5D responses shows that SS (physician) has lower agreement with patient responses along the *Mobility* dimension than does PH (non-physician) likely due to SS classifying fewer patients as having mobility problems (Pneg = .59 vs. .71). The agreement along other dimensions is consistently in the "moderate" category. The negative agreement on the *Pain* dimension is higher that positive agreement indicating that both raters found more evidence for pain in the EMR than self-reported by patients. This is reversed on the other dimensions.

Table 4 Agreement between HRQOL dimensions of EQ5D and manual classification of clinical notes

	ICC (Ppos)(Pneg)		
	SS vs. EQ5D	**PH vs. EQ5D**	**SS vs. PH**
Pain	.54 (.51)(.84)	.57 (.48)(.85)	.70 (.55)(.88)
Mobility	.51 (.72)(.59)	.63 (.73)(.71)	.48 (.67)(.58)
Usual act	.55 (.72)(.63)	.55 (.72)(.64)	.70 (.73)(.68)
Depress.	.58 (.76)(.58)	.63 (.74)(.67)	.74 (.80)(.71)

Several key observations can be made based on the results of this study. First, both manual and automatic classification approaches resulted in "moderate" agreement with patient responses. This finding indicates that care providers' assessment of the patient's quality-of-life characteristics differs from the patients' own perceptions. However, our results also indicate that using physician's clinical notes as a surrogate measure of patient's HRQOL works reasonably well for the *Pain* dimension.

Table 5 Agreement between HRQOL dimensions of the EQ5D index and SVM classification of clinical notes

	Feature extraction Method					
	Bag-of-words			Bag-of-concepts		
	Feature selection Method					
Dimension	NFS	IG	CFS	NFS	IG	CFS
Pain			112*			112
Kappa	.17	.45	.48	.25	.63	.36
Ppos	.35	.62	.63	.44	.74	.46
Pneg	.80	.83	.84	.80	.88	.87
Mobility			106			123
Kappa	.22	.36	.47	.26	.48	.55
Ppos	.73	.76	.81	.73	.80	.84
Pneg	.48	.59	.65	.52	.67	.70
Usual act.			114			86
Kappa	.08	.53	.33	.32	.49	.55
Ppos	.95	.97	.82	.79	.84	.87
Pneg	.12	.56	.49	.51	.65	.68
Depression			64			123
Kappa	.13	.34	.40	.07	.48	.47
Ppos	.75	.79	.83	.70	.84	.84
Pneg	.36	.54	.55	.37	.63	.62
Mean Kappa	.15	.42	.42	.23	.52	.48

* Number of predictive features selected by the CFS algorithm

Both EQ5D and SF6D instruments had consistently better agreement along this dimension with the automatic classifiers trained on clinical notes. This may be due to the fact that physical pain is a highly salient characteristic for the patient and probably tends to be reported and assessed during an office visit more frequently and in more detail by the clinician than other aspects of patient functioning. The two instruments seem to differ with respect to the *Depression/Mental* dimensions, where SF6D has a much better agreement with automatic classifiers than EQ5D. This may be due to the more detailed nature of the SF-36 questionnaire that incorporates questions that may be more sensitive in eliciting patient's anxiety or depression.

We do not assume that it is optimal to have perfect agreement between patients' responses to functional assessment instruments and the clinician's subjective assessment of characteristics such as pain, depression and patient's social and usual activities. While clinician notes may not contain comprehensive information on the patient's functional status, we use the notes as a starting point in this research to explore patient-physician communication regarding HRQL, its documentation and use of NLP to predict patient outcomes.

Another future direction is to examine functional status stability over time. Our current methodology relies on pooling the information from a number of patient visits prior to a certain point in time (index visit).

Conclusions

Our results indicate that using the bag-of-concepts approach to feature extraction from the text of clinical reports that relies on the Metamap for concept identification is advantageous over the baseline bag-of-words technique. We also find that, using a cascade of the IG and the CFS feature selection produces better results with the exception of the EQ5D pain dimension. Our preliminary findings indicate that clinician's notes may be used to develop automated surrogate measures of HRQOL status, but this requires further investigation.

Acknowledgements

This work was made possible by Grant #1 UL1 RR024150* from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH), and the NIH Roadmap for Medical Research.

References

1. Pakhomov S, Hanson PL, Bjornsen SS, Smith SA. Quality Performance Measurement Using the Text of Electronic Medical Records. Medical Decisions Making 2008;(in press).

2. Melton LJ, 3rd. History of the Rochester Epidemiology Project. Mayo Clinic Proc 1996;71(3):266-74.

3. Hanmer J, Lawrence WF, Anderson JP, Kaplan RM, Fryback DG. Report of nationally representative values for the noninstitutionalized US adult population for 7 health-related quality-of-life scores.[see comment]. Medical Decision Making 2006;26(4):391-400.

4. Hawthorne G, Osborne RH, Taylor A, Sansoni J. The SF36 Version 2: critical analyses of population weights, scoring algorithms and population norms. Quality of Life Research 2007;16(4):661-73.

5. Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. Health Economics 2004;13:873-84.

6. Schmittdiel J, Vijan S, Fireman B, Lafata JE, Oestreicher N, Selby JV. Predicted Quality-Adjusted Life Years as a Composite Measure of the Clnical Value of Diabetes Risk Factor Control. Medical Care 2007;45(4):315-21.

7. Shaw JW, Johnson JA, Coons SJ. US valuation of the EQ-5D health states: development and testing of the D1 valuation model. Medical Care 2005;43(3):203-20.

8. Tsuchiya A, Brazier J, Roberts J. Comparison of valuation methods used to generate the EQ-5D and the SF-6D value sets. Journal of Health Economics 2006;25(2):334-46.

9. Dolin RH, Alschuler L, Boyer S, et al. HL7 Clinical Document Architecture, Release 2. J Am Med Inform Assoc 2006;13(1):30-9.

10. Cohen AM. An effective general purpose approach for automated biomedical document classification. AMIA Annu Symp Proc 2006:161-5.

11. Xu H, Markatou M, Dimova R, Liu H, Friedman C. Machine learning and word sense disambiguation in the biomedical domain: design and evaluation issues. BMC Bioinformatics 2006;7:334.

12. Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis C. Text categorization models for high-quality article retrieval in internal medicine. J Am Med Inform Assoc 2005;12(2):207-16.

13. Hiissa M, Pahikkala T, Suominen H, et al. Towards automated classification of intensive care nursing narratives. Stud Health Technol Inform 2006;124:789-94.

14. Joshi M, Pakhomov SV, Pedersen T, Chute CG. A comparative study of supervised learning as applied to acronym expansion in clinical reports. AMIA Annu Symp Proc 2006:399-403.

15. Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. 2nd ed. San Francisco: Elsevier; 2005.

16. Lewis DP, Jebara T, Noble WS. Support vector machine learning from heterogeneous data: an empirical analysis using protein sequence and structure. Bioinformatics 2006;22(22):2753-60.

17. Noble WS. What is a support vector machine? Nat Biotechnol 2006;24(12):1565-7.

18. Platt J. Fast Training of Support Vector Machines using Sequential Minimal Optimization. In: Schoelkopf B, Burges C, Smola A, eds. Advances in Kernel Methods - Support Vector Learning. Boston, MA: MIT Press; 1998.

19. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings / AMIA Annual Symposium 2001:17-21.

20. Browne AC, Divita G, Aronson AR, McCray AT. UMLS language and vocabulary tools. AMIA Annu Symp Proc 2003:798.

21. Rindflesch TC, Tanabe L, Weinstein JN, Hunter L. EDGAR: Extraction of Drugs, Genes and Relations from the Biomedical Literature. In proceedings of the Pacific Symposium on Biocomputing 2000:517-28.

22. Mark AH, Geoffrey H. Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. IEEE Trans on Knowl and Data Eng 2003;15(6):1437-47.

23. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. Journal of clinical epidemiology 1990;43(6):543-9.

24. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. Journal of clinical epidemiology 1990;43(6):551-8.

25. St Sauver JL, Hagen PT, Cha SS, et al. Agreement between patient reports of cardiovascular disease and patient medical records. Mayo Clinic proceedings 2005;80(2):203-10.

26. Shrout PE, Fleiss JL. Intraclass Correlations: Uses in Assessing Rater Reliability. Psychological Bulletin 1979;86(2):420-8.