Classifying Encounter Notes in the Primary Care Patient Record

Thomas Brox Røst and **Øystein Nytrø** and **Anders Grimsmo**[1]

Abstract. The ability to automate the assignment of primary care medical diagnoses from free-text holds many interesting possibilities. We have collected a dataset of free-text clinical encounter notes and their corresponding manually coded diagnoses and used it to built a document classifier. Classifying a test set of 2,000 random encounter notes yielded a coding accuracy rate of 49.7 %. Automated coding of primary care encounter notes is a novel application area, and though imperfect our method proves interesting enough to warrant further research.

1 Introduction

In this study we attempt to classify primary care clinical encounter notes into their corresponding diagnoses. We do so by learning document classifiers from a manually coded dataset collected from a Norwegian primary care center. Research have shown that the manual diagnosis coding of primary care encounter notes tend to be of high quality [20]. This, coupled with the the size of the dataset, makes the application area interesting from an information retrieval and document classification point of view. In the long term, being able to infer diagnoses from written text might prove useful in e.g. detection of incorrect diagnoses and improving electronic patient record systems. We consider this study as an initial exploration into applying proven document classification techniques onto a novel application area.

The electronic patient record (EPR) has gradually attained widespread usage in primary care. In Norway, more than 90 % of primary care physicians are routinely using computer-based patient-record systems [3] and many have been doing so for more than 15 years. A typical feature of most commercial EPR systems in use today is that the encounter note, which is the main documentation of the doctor-patient consultation, is written as free-text narrative. There are perfectly practical reasons for this: Unstructured free-text is easy to write and represents the traditional way of documenting patient treatment. However, this makes the information within less suitable for automated processing and thereby keeps the EPR from fulfilling its full potential as a useful tool for both research and clinical practice. Attempts have been made to create EPRs that impose varying degrees of structure on the clinical narrative, but with limited success so far.

To alleviate this problem, many researchers have attempted to use natural language processing (NLP), text classification and text mining techniques on clinical narrative. Some NLP systems have proven very useful in a number of clearly defined domains, such as detection of bacterial pneumonia from chest X-ray reports [4], finding adverse drug events in outpatient medical records [10] and discharge summaries [19], and identifying suspicious findings in mammogram reports [12]. A common feature of such systems is that they restrict themselves to a narrow clinical domain with a clearly defined vocabulary and a limited form of discourse, such as one would find in specialized hospital reports. Our long-term goal is to draw on research from these areas and explore the usefulness of similar techniques on the primary care patient record. However, the lack of empirical knowledge on the content in primary care documentation raises the need for preliminary investigations on the narrative structure found therein. This initial study attempts to use supervised document classfication to explore if there is a correspondence between the diagnosis and the documented encounter. Besides from the previously mentioned possible benefits of automated coding, a secondary purpose is to learn more about the informational value and underlying documentational patterns in primary care encounter notes.

2 Background

Among the characteristic features of primary care encounter notes are sparseness, brevity, heavy use of abbreviations and many spelling mistakes. The notes are normally written during the consultation by the treating physician, this in contrast with hospital patient records which are usually dictated by the physician and then transcribed by a secretary. A typical encounter note might look something like this:

> Inflamed wounds over the entire body. Was treated w/ apocillin and fucidin cream 1 mth. ago. Still using fucidin. Taking sample for bact. Beginning tmnt. with bactroban. Call in 1 week for test results[2].

To classify such notes we rely on the presence of manually coded diagnosis codes. The use of clinical codes in primary care is common in the United Kingdom, the Netherlands, and Norway [16]. The motivation for coding is both for reimbursement and statistical purposes. In our experimental dataset the notes are coded according to the ICPC-2 coding system. ICPC-2 is the second edition of the International Classification of Primary Care, a coding system which purpose is to provide a classification that reflects the particular needs and aspects of primary care [11]. Using a single ICPC code, each health care encounter can be classified so that both the reasons for encounter, diagnoses or problems, and process of care are evident. Together, these elements make out the core constituent parts of the health care encounter in primary care. Moreover, one or more encounters associated with the same health problem or disease form an episode of care [9].

[1] Department of Computer and Information Science and The Norwegian EHR Research Centre, Norwegian University of Science and Technology, Trondheim, Norway. Emails: {brox, nytroe}@idi.ntnu.no, anders.grimsmo@medisin.ntnu.no

[2] Translated from the Norwegian

ICPC-2 follows a bi-axial structure with 17 chapters along one axis and 7 components along the other. The chapters are single-letter representations of body systems (Table 1) while the components are two-digit numeric values (Table 2). As an example, "R02" is the ICPC code for shortness of breath.

Table 1. ICPC chapter codes.

Chapter code	Description
A	General and unspecified
B	Blood, blood-forming organs and immune mechanism
D	Digestive
F	Eye
H	Ear
K	Circulatory
L	Musculoskeletal
N	Neurological
P	Psychological
R	Respiratory
S	Skin
T	Endocrine, metabolic and nutritional
U	Urological
W	Pregnancy, child-bearing, family planning
X	Female genital
Y	Male genital
Z	Social problems

Table 2. ICPC component codes.

Number	Range	Description
1	01-29	Complaint and symptom component
2	30-49	Diagnostic, screening, and preventive component
3	50-59	Medication, treatment, procedures component
4	60-61	Test results component
5	62-63	Administrative component
6	64-69	Referrals and other reasons for encounter
7	70-99	Diagnosis/disease component

There are several examples of attempts to automate the coding of diagnoses [5, 15, 18, 21, 23], all of which concern themselves with the alternative ICD code. ICD is a more complex code than ICPC and is more suited for specialized usage in hospitals. March [18] describes the use of Bayesian learning to achieve automated ICD coding of discharge diagnoses. Franz [5] compares coding methods with and without the use of an underlying lexicon and concludes that lexicon-based methods perform no better than lexicon-free methods, unless one adds conceptual knowledge. Larkey [15] found that using a combination of different classifiers yielded improved automatic assignment of ICD codes. There is a practical purpose to automated ICD coding: ICD is a more complex code than ICPC and accordingly manual ICD encoding takes up a lot of time. There have also been other approaches towards automated coding of clinical text. Hersh [8] attempted to predict trauma registry procedure codes from emergency room dictations. Aronow [2] classified encounter notes in order to find acute exacerbations of asthma and radiology reports for certain findings, this through the use of Bayesian inference networks and the ID3 decision tree algorithm. Document classification and IR has been applied in other medical domains as well, such as clustering of medical paper abstracts [17].

Examples of automated ICPC coding are harder to come by. Letrilliart [16] describes a string matching system that assigns ICPC codes from free-text sentences containing hospital referral reasons, based on a manually created look-up table. We have not found examples of similar attempts at automated ICPC classification in the literature.

As for classification techniques, this study uses support vector machines (SVM). SVMs have proved useful and have shown good general performance for text classification tasks [13] when compared with other classifiers. Our goal for this study is not to compare classification methods; this will be explored further in future work.

3 Methods and Data

We have collected a dataset from a medium-sized general practice office in Norway. The data consists of encounter notes for a total of 10,859 patients in the period from 1992 to 2004. All in all, there are 482,902 unique encounters. The Norwegian Health Personnel Act [1] requires that caregivers provide "relevant and necessary information about the patient and about the health care" in the patient record. In practice, this manifests itself as a combination of structured and unstructured information about the encounter. Information such as personal details about the patient, prescriptions, laboratory results, medical certificates and diagnosis codes is typically available in structured format, while encounter notes, referrals and discharge notes comes in the form of unstructured free-text. For the purposes of this paper, we have only considered the encounter notes and the accompanying ICPC-2 diagnosis code.

A known source of noise is that a minority of the notes are likely to be written in Danish or *nynorsk* (literally "New Norwegian") rather than standard Norwegian (*bokmål*). There are also more than 20 different authors, so there may be differences in documentational style as well. Interns fresh out of medical school may for example be more inclined to document more thoroughly than an experienced physician.

The dataset has been automatically anonymized using a custom-built anonymization tool [22]. Each word or token is controlled against a database of words that are known to be insensitive and a set of rules that deal with alphanumeric patterns such as medication doses, date ranges, and laboratory test values. Sensitive tokens are replaced with a general identifier or an identifier that shows the type of token that was replaced.

Each encounter will typically consist of a written note of highly variable length and zero or more accompanying ICPC codes. 287,868 of the available encounters have one or more ICPC codes (Table 3).

Table 3. Number of ICPC codes per encounter.

Number of ICPC codes	Number of encounters
1	235,860
2	44,651
3	6,037
$\geq$4	1,320

There are some notable differences in terms of code use between hospital and primary care settings. Larkey [15] describes a test set of discharge summaries with a mean of 4.43 ICD-9 codes per document, while Nilsson [20] notes that a set of Swedish general practice patient records has a mean of 1.1 ICD-10 codes per record. While there may be regional and cultural differences with respect to coding practice, the latter corresponds with our findings of 1.2 ICPC-2 codes per note (Table 3).

Since we concern ourselves with the relation between the encounter note and the ICPC code, we discard all encounters with more

than one code in order to avoid ambiguity in the training data. Of the 235,860 encounters that are left, 175,167 have an accompanying encounter note.

The use of ICPC codes as classification bins for encounter notes is essentially a multi-class classification problem. Since there are 726 distinct ICPC codes it becomes practical to reduce the class dimensionality. We choose to group codes according to their chapter value, so that we are left with the 17 single-letter body codes as classes.

When grouping encounter notes by their ICPC chapter value we note that there is a varying degree of verbosity. The use of sparse encounter notes is often common in primary care, for instance when renewing recurring prescriptions. To determine average note verbosity for each ICPC chapter, all relevant encounter notes are tokenized. After removing stop words, whitespace and other noisy elements, the average length and standard deviation is calculated (Table 4).

Table 4. Average note length by ICPC chapter.

Chapter	Avg. No words	St. dev.	Samples
N (Neurological)	40	33.2	5,637
D (Digestive)	39	30.0	11,386
Z (Social)	36	35.1	570
X (Female genital)	36	27.1	6,244
P (Psychological)	32	35.6	9,939
A (General)	32	28.9	12,052
Y (Male genital)	31	24.9	1,993
F (Eye)	31	23.5	4,998
L (Musculoskeletal)	29	26.8	36,493
R (Respiratory)	28	21.8	22,846
K (Circulatory)	27	25.6	21,089
H (Ear)	27	21.3	5,526
W (Pregnancy)	26	24.5	5,614
U (Urological)	26	25.2	4,502
T (Endocrine)	26	22.4	5,498
S (Skin)	26	20.3	18,432
N/A	23	20.6	6,545
B (Blood)	22	23.3	2,348

We note that Larkey's discharge summaries [15] has a mean length of 633 words, which is more than an order magnitude higher than our notes. Notwithstanding cultural and institutional differences, this highlights how hospital discharge summaries usually provide a more self-contained description of the patient and his ailments. In the Norwegian health care system the patient will typically use just one primary care physician who acts as a gatekeeper for specialized hospital care when necessary. Accordingly descriptions of the patient's state may span several encounter notes in the primary care patient record.

Since many classification techniques, including support vector machines (SVM), are restricted to dealing with binary classification tasks, we have to reduce our multi-class classification task into a set of binary tasks. For each pair of classes $(i, j) : i, j \in \{A, B, \ldots, Z\}$ where $i, j = 1 \ldots c, j \neq i$ we create a two-class classifier $< i, j >$. If c is the number of classes, we end up with $c(c-1)$ binary classifiers, or $17 \times 16 = 272$ in this case. This technique is known as double round robin classification [6]. The classifier $< i, j >$ will then solely consist of training examples from encounter notes with ICPC chapter codes i and j. To determine the final predicted class of any given note we feed it through each classifier and record the result. The class that receives the highest number of predictions is chosen to be the most likely one. In case of ties we choose the class with the highest number of occurrences in the training set, or, as a last resort, pick one at random. To build and run the classifiers we used the SVM-Light[3] toolkit.

We use word and phrase frequencies as the base component when constructing feature vectors for the classifiers. If we were to rely on single words alone we would lose some contextual information [8], so frequency counts are performed on all unigrams, bigrams and trigrams in the encounter note, excluding stop words. The occurrence of an n-gram is recorded as a *true* value in the feature vector. While n-grams may be a simplistic way of representing context, it still allows us to catch phrases and turns of words that may have discerning qualities.

As is common with word-based feature vectors, it is useful to apply some dimension-reducing technique to limit the size of the vector. The challenge lies in pruning those features that are the most inconsequential to the classifier's predictive qualities. For this experiment we adapt a technique described in [14]. For each classifier the frequency of all unigrams, bigrams and trigrams occurring in all training notes for both classes are counted. If an n-gram occurs in more than 7.5 % of either the true or the false class notes it is tagged as a likely candidate for inclusion. All candidates are then ranked according to their true class frequency to false class frequency ratio. Finally the top 100 candidates are chosen as the most relevant features. As an example, Table 5 shows the 20 first selected features from the F (Eye) versus P (Psychological) classifier.

Table 5. F versus P classifier, 20 most relevant features.

Original n-gram	Appr. English translation	Comment
kloramf	chloramph	Abbreviation
cornea	cornea	
øyelokk	eyelid	
rusk	dust	
hø øye	right eye	Abbreviation
kloramfenikol	chloramphenicol	
rdt	red	
ve øye	left eye	Abbreviation
øye	eye	
øyet	the eye	
injeksjon	injection	
puss	pus	
øyne	eyes	
hø	right	Abbreviation
ve	left	Abbreviation
begge	both	Abbreviation
ved us	after examination	Abbreviation
us	examination	Abbreviation
lett	easily	
ser	sees	

2.000 notes were selected at random from the 175.167 available notes to be used as a test set; the remaining notes were used to train the classifiers. As seen from Table 4, this implies that the amount of training data available for each classifier will differ.

4 Results

Table 6 shows the results from attempting to classify the 2.000 test cases. A total of 994 cases were classified correctly, giving an overall accuracy rate of 49.7 %. As a comparison, guessing for the most frequent chapter code (L) all the time will yield an accuracy of 20.8 %. The displayed results are from a single test run.

[3] http://svmlight.joachims.org/

Table 6. Predicted classes of 2,000 notes in test set.

Correct ICPC chapter	Predicted ICPC chapter																		
	A	B	D	F	H	K	L	N	P	R	S	T	U	W	X	Y	Z	Sum	% correct
A	**13**	0	10	0	0	13	71	0	3	25	12	0	0	0	2	0	0	149	8.7 %
B	0	**0**	0	0	0	1	25	0	0	6	0	0	0	0	0	0	0	32	0.0 %
D	1	0	**64**	0	0	1	47	0	0	4	9	0	0	0	1	0	0	127	50.3 %
F	0	0	0	**19**	0	1	30	1	0	5	2	0	0	0	0	0	0	58	32.7 %
H	0	0	0	0	**16**	2	29	0	0	10	4	0	0	0	1	0	0	62	25.8 %
K	0	0	3	0	0	**158**	56	0	0	5	0	0	0	0	1	0	0	223	70.8 %
L	0	0	3	0	0	5	**348**	1	0	5	9	0	1	0	1	0	0	373	93.2 %
N	2	0	2	0	0	9	42	**4**	3	1	0	0	0	0	3	0	0	66	6.0 %
P	1	0	2	0	0	5	93	0	**33**	4	0	0	0	0	3	0	0	141	23.4 %
R	3	0	3	0	0	5	73	0	0	**170**	2	0	0	0	2	0	0	258	65.8 %
S	0	0	2	0	3	2	84	0	1	3	**128**	0	0	0	0	0	0	223	57.3 %
T	1	0	2	0	0	8	30	1	5	2	0	**2**	0	0	2	0	0	53	3.7 %
U	0	0	0	0	0	2	31	0	5	1	2	0	**1**	0	0	0	0	42	2.3 %
W	0	0	0	0	0	7	56	0	1	0	0	0	0	**15**	4	0	0	83	18.0 %
X	0	0	6	0	0	8	45	0	1	3	1	0	0	3	**23**	0	0	90	25.5 %
Y	0	0	1	0	0	2	14	0	1	0	0	0	1	0	0	**0**	0	19	0.0 %
Z	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	**0**	1	0.0 %

5 Discussion and Future Work

When considering the results, we must bear in mind that they are from a single run. To verify their validity they should be averaged over several test runs of independent samples.

Even though the accuracy varies a lot for the individual chapters, the results are still quite promising. The most notable feature is how the L (Muscoloskeletal) class appears to soak up the majority of the misclassified cases. We are not sure why this is happening. The L group constitutes the largest group in the training set, followed by the R, K and S groups. When attempting to perform the same classification task without the L cases the S group became the major misclassification bin, but in a less dramatic fashion; the overall accuracy rate rose to 57.5 %. In general, our naive, largely domain-ignorant approach granted results that are interesting enough to legitimate further work in this area.

There are several possible approaches to approving the predictive quality of the classifier. We made no attempts to normalize the vocabulary in the training data. Techniques such as stemming or mapping terms to a common controlled vocabulary would reduce the number of relevant features. This would also involve dealing with common misspellings [7] and dialect terms, both of which are quite common in our dataset. Wilcox [24] notes that the use of expert knowledge can provide a significant boost to medical text report classifiers. It would also be worth investigating if the use of accompanying information from the EPR, such as lab results and prescriptions, can help improve classification quality. Another possible approach is to view the encounter note in its longitudinal context by also considering notes from previous (and following) encounters.

We made no efforts to control the amount of noise in the classifiers or to screen the notes in the test data set. Very short notes and notes with non-standard language use were not discarded. Also, the influence of n-gram feature threshold selection on the quality of the results could have been evaluated. Similarly, the effect of using additional parameters such as average note length and n-gram partial coincidence would have been worth investigating.

The a priori anonymization could also influence the results. Since the anonymization tool only allows known non-sensitive words, it is likely that special and unusual words are lost. Such words may have a higher predictive effect than more common words. Comparing the classifier on a non-anonymized dataset could possibly indicate how much of destructive effect that is incurred due to anonymization.

The choice of ICPC chapter codes as class indicators is not necessarily a natural choice. Indeed, this may be seen as a simplification of the problem of diagnosis prediction. Alternatives include grouping according to ICPC component codes or, as a natural follow-up, attempting to classify into the full ICPC codeset of 726 different codes.

ACKNOWLEDGEMENTS

Thanks go to Amund Tveit, Ole Edsberg, Inger Dybdahl Sørby and Gisle Bjørndal Tveit for comments and suggestions.

REFERENCES

[1] Act of 18 may 2001 no. 24 on personal health data filing systems and the processing of personal health data, 2004.04.12 2001.

[2] D. B. Aronow, S. Soderland, J. M. Ponte, F. Feng, W. B. Croft, and W. G. Lehnert, 'Automated classification of encounter notes in a computer based medical record', *Medinfo*, **8 Pt 1**, 8–12, (1995).

[3] Elisabeth Bayegan, *Knowledge Representation for Relevance Ranking of Patient-Record Contents in Primary-Care Situations*, Ph.D. dissertation, Norwegian University of Science and Technology (NTNU), 2002.

[4] M. Fiszman, W. W. Chapman, D. Aronsky, R. S. Evans, and P. J. Haug, 'Automatic detection of acute bacterial pneumonia from chest x-ray reports', *J Am Med Inform Assoc*, **7**(6), 593–604, (2000). Evaluation Studies Journal Article.

[5] Pius Franz, Albrecht Zaiss, Stefan Schulz, Udo Hahn, and Rdiger Klar, 'Automated coding of diagnoses - three methods compared', in *Proceedings of the Annual Symposium of the American Society for Medical Informatics (AMIA)*, Los Angeles, CA, USA, (2000).

[6] Johannes Frnkranz, 'Round robin classification', *J. Mach. Learn. Res.*, **2**, 721–47, (2002).

[7] W. R. Hersh, E. M. Campbell, and S. E. Malveau, 'Assessing the feasibility of large-scale natural language processing in a corpus of ordinary medical records: a lexical analysis', *Proc AMIA Annu Fall Symp*, 580–4, (1997).

[8] W. R. Hersh, T. K. Leen, P. S. Rehfuss, and S. Malveau, 'Automatic prediction of trauma registry procedure codes from emergency room dictations', *Medinfo*, **9 Pt 1**, 665–9, (1998).
[9] I. M. Hofmans-Okkes and H. Lamberts, 'The international classification of primary care (icpc): new applications in research and computer-based patient records in family practice', *Fam Pract*, **13**(3), 294–302, (1996).
[10] B. Honigman, P. Light, R. M. Pulling, and D. W. Bates, 'A computerized method for identifying incidents associated with adverse drug events in outpatients', *Int J Med Inform*, **61**(1), 21–32, (2001). Journal Article.
[11] WONCA International, *ICPC-2: International Classification of Primary Care*, Oxford Medical Publications, 2 edn., 1998.
[12] N. L. Jain and C. Friedman, 'Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports', *Proc AMIA Annu Fall Symp*, 829–33, (1997).
[13] Thorsten Joachims, 'Text categorization with suport vector machines: Learning with many relevant features', in *ECML '98: Proceedings of the 10th European Conference on Machine Learning*, pp. 137–142, London, UK, (1998). Springer-Verlag.
[14] Andries Kruger, C. Lee Giles, Frans Coetzee, Eric Glover, Gary Flake, Steve Lawrence, and Cristian Omlin, 'Deadliner: Building a new niche search engine', in *Ninth International Conference on Information and Knowledge Management, CIKM 2000*, Washington, DC, (2000).
[15] Leah S. Larkey and W. Bruce Croft, 'Combining classifiers in text categorization', in *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 289–97, Zurich, Switzerland, (1996). ACM Press.
[16] L. Letrilliart, C. Viboud, P. Y. Boelle, and A. Flahault, 'Automatic coding of reasons for hospital referral from general medicine free-text reports', *Proc AMIA Symp*, 487–91, (2000).
[17] Pavel Makagonov, Mikhail Alexandrov, and Alexander Gelbukh, 'Clustering abstracts instead of full texts', *Lecture Notes in Computer Science*, **3206**, 129–35, (2004).
[18] Alan D. March, Eitel J. M. Laura, and Jorge Lantos, 'Automated icd9-cm coding employing bayesian machine learning: a preliminary exploration', in *Simposio de Informtica y Salud 2004*, (2004).
[19] G. B. Melton and G. Hripcsak, 'Automated detection of adverse events using natural language processing of discharge summaries', *J Am Med Inform Assoc*, **12**(4), 448–57, (2005).
[20] G. Nilsson, H. Ahlfeldt, and L. E. Strender, 'Textual content, health problems and diagnostic codes in electronic patient records in general practice', *Scand J Prim Health Care*, **21**(1), 33–6, (2003). Journal Article.
[21] Y. Satomura and M. B. do Amaral, 'Automated diagnostic indexing by natural language processing', *Med Inform (Lond)*, **17**(3), 149–63, (1992).
[22] Amund Tveit, Ole Edsberg, Thomas Brox Røst, Arild Faxvaag, Øystein Nytrø, Torbjørn Nordgård, Martin Thorsen Ranang, and Anders Grimsmo, 'Anonymization of general practioner's patient records', in *Proceedings of the HelsIT'04 Conference*, Trondheim, Norway, (2004).
[23] Rodrigo F. Vale, Berthier A. Ribeiro-Neto, Luciano R.S. de Lima, Alberto H.F. Laender, and Hermes R.F. Junior, 'Improving text retrieval in medical collections through automatic categorization', *Lecture Notes in Computer Science*, **2857**, 197–210, (2003).
[24] A. B. Wilcox and G. Hripcsak, 'The role of domain knowledge in automating medical text report classification', *J Am Med Inform Assoc*, **10**(4), 330–8, (2003).