MEDINFO 2007
HISA Ltd Australia
© 2007 The authors. All rights reserved.

Collecting Presenting Problems From Patient Records by Text Mining

Anders Thurin

Sahlgrenska University Hospital, Göteborg, Sweden

Abstract

The presenting problem in emergency care largely determines the direction of activities and thus the workload of parts of the hospital. To determine the mixture of presenting problems in the emergency department of a teaching hospital we analysed text in patient records, using a set of regular excpressions to extract the patients' presenting problem/reason for encounter. Such a problem could be extracted in about 84% of cases, and suggests that such methods can be useful for creating statistics. Among anticoagulated patients, bleeding related problems were overrepresented, suggesting that this method could also be used to find adverse drug effects.

Keywords:

natural language processing, data mining, reason for encounter, adverse drug reactions, anticoagulant therapy

Introduction

A majority of the patients treated at our hospital present as non-planned emergency visits and are initially admitted through the emergency ward. When a patient is discharged the final diagnosis is reported for national statistics, but no information is collected routinely on the reasons for encounter. The distribution of patient problems in the emergency ward to a determines a large part of the workload and activities at the hospital. Manual coding of patients' reasons for encounter is time consuming, and no optimal coding system is available.

The aim of this study was to analyse the distribution of reasons for encounter in general as well as among patients treated with the anticoagulant drug warfarin. Data extracted this way could be aggregated to describe the case mix of actual patients, and help in planning the work and activities at the hospital. One hypothesis was that conditions related to bleeding would be more common with anticoagulation.

Materials and methods

Materials

All records from 1994-2005 of the database Melior[1], Departments of Surgery and Internal medicine, Östra Sjukhuset, Sahlgrenska University Hospital, Göteborg, Sweden, were analysed. We have no access to the database tables containing patient identity, which helps preserve patient integrity. The method of data collection has been approved by the local ethics committee and those responsible for security at the hospital.

Text in the patient record is sorted in the database under different record headings, and for this study only text under the heading "besöksorsak" (reason for encounter) from the physician emergency note was analysed. This text can consist of a single word, like "buksmärtor" (abdominal pains), or it can be made up of anything from a few words to several sentences, e.g. "patienten söker på grund av illamående och buksmärtor" (the patient seeks help because of nausea and abdominal pains). In addition to the reason for encounter text under this heading often describes other conditions of patients' admission, such as from where patient is referred (other hospital, GP, other clinic), or how patient comes to the emergency ward (by ambulance/brought in by relative…).

Methods

Text paragraphs were collected from the database by SQL queries, giving a relatively homogeneous set of text paragraphs, most of which contained a reason for encounter. The average length of each text line in medical and surgical notes were 38.7 and 38.3 characters, respectively. The longest text string was 1444 characters and the shortest 1 character. The result set was saved in a text file consisting of 49918 lines from surgery, 87037 from internal medicine, with text from one medical record on each line. Similar text files of 1035 lines from surgery, 8791 from internal medicine was produced with the additional criterion that the medical record somewhere contained the string "Waran", the trade name for Warfarin in Sweden.

The search method was developed using only the surgery material, and later applied to the internal medicine text. First analysis was made using a concordance (a list of all the words in the text presented with their closest context) was created from the larger text file using the software package Concordance 3.0[2], which count the number of instances of every unique word, and allows concordance lists to be sorted by the context that appears around the word - the word before or after, two words before or after, and so on.

1 Siemens Health Solutions

2 R.J.C Watts. http://www.concordancesoftware.co.uk/

S0rting the words by number of appearances made it possible to see which reasons for encounter were the most common. By looking at the concordance of a common reason for encounter with the context sorted in different ways it was easy to see in which surroundings the word often appears. Patterns in the text that were often associated with a reason for encounter were determined by manual analysis of concordances, and guided by these identified patterns, customized Perl[3] scripts could identify and count phrases matching reasons for encounter according to the patterns determined in the concordance analysis. The method of doing this was divided into two steps. In the first step, a Perl script reads through the text file one line at a time, trying to match each line with a pattern such as:

- a single word, like buksmärtor (abdominal pains) or skalltrauma (head trauma).
- two or a few words, or comma separated lists of such word groups, like commotio cerebri or svårighet att gå, buksmärtor (difficulties to walk, abdominal pains)
- a comma separated lists of single words, like buksmärta, illamående, diarreer (abdominal pains, nausea, diarrhea) or hematemes, melena samt buksmärtor (hematemes, melena and abdominal pains).

Table 1 - Reasons for encounter in surgery

Surgery - total		*Occurrences*	
Reason for encounter		*n*	*%*
Buksmärtor	abdominal pains	9036	18,0
Buksmärta	abdominal pain	4224	8,43
Kräkningar	vomiting	1455	2,90
Misshandel	abuse	953	1,90
Sårskada	wound	834	1,66
Illamående	nausea	813	1,62
Feber	fever	712	1,42
Falltrauma	fall	609	1,22
Trafikolycka	traffic accident	585	1,17
Diarreer	diarrhea	512	1,02
Skalltrauma	head injury	412	0,82
Melena	melena	409	0,82
Hematemes	hematemesis	402	0,80
trauma skalle	head trauma	383	0,76
Diarre	diarrhea	344	0,69
Förstoppning	obstipation	268	0,54
Obstipation	obstipation	264	0,53
epigastralgier	epigastralgia	257	0,51
Urinstämma	urinary obstruction	209	0,42
trauma thorax	thorax trauma	181	0,36

- longer text strings where keywords or keyphrases indicate the reason for encounter, like ...frågeställning om appendicit (suspected appendicitis) or ...söker pga högersidig buksmärta, illamående och kräkningar (seeking care due to abdominal pain, nausea and vomiting). In these cases only the rest of the sentence after the keyphrase (here frågeställning om and söker pga) is regarded as the reason for encounter, even if there is more text after that sentence.

In the second step another Perl script reads through the hashtable to further split and analyze the phrases identified in step 1, often containing long strings of text or comma separated lists of words, by pattern matching.

Some examples of patterns handled in step 2 are:

- Words in lists and enumerations separated by commas, full stops and conjunctions are regarded as separate reasons for encounter, like buksmärta, illamående, diarreer (abdominal pain, nausea, diarrhea) which result in three reasons for encounter.
- Lists or enumerations of word groups consisting of a few words separated by commas, full stops and conjuctions are regarded as separate reasons for encounter, like svårighet att gå, högersidiga buksmärtor (difficulties to walk, abdominal pains on the right side). Here svårighet att gå is one reason for encounter, högersidiga buksmärtor is another.

Results

The Perl scripts were able in the surgery material to identify a reason for encounter according to the patterns in 42961 of the 49918 lines in the text file (86.1 %). In all, 50093 occurrences of any of 10459 different reasons for encounter were found. 406 reasons for encounter were identified 10 times or more, accounting for 35934 occurrences (72% of all). 8283 reasons for encounter only appear once (16% of occurrences, 79% of types) some of which may be e.g. spelling errors or in need of further processing. The most common reasons for encounter consist of one or two words, like illamående (nausea) or trauma skalle (head trauma). 3387 (32% of types) of the reasons for encounter consist of one or two words, and most of these seem to be correct and relevant.

Table 2 - Reasons for encounter in surgery among warfarin patients

Surgery - Waran		*Occurrences*	
Reason for encounter		*n*	*%*
buksmärtor	abdominal pains	158	14,74
buksmärta	abdominal pain	58	5,41
melena	melena	27	2,52
falltrauma	fall	25	2,33
illamående	nausea	23	2,15
kräkningar	vomiting	23	2,15
hematemes	hematemesis	19	1,77
feber	fever	18	1,68
hematuri	hematuria	14	1,31
gastrointestinal blödning	GI bleeding	13	1,21
rektalblödning	rectal bleeding	12	1,12
svart avföring	black stools	10	0,93
sårskada	wound	10	0,93
diarre	diarrhea	10	0,93
urinstämma	urinary obstruction	9	0,84
ikterus	icterus	8	0,75
trauma skalle	head trauma	8	0,75
anemi	anemia	8	0,75
blödning per rektum	rectal bleeding	7	0,65

3 ActivePerl 5.8 http://www.activestate.com

In the Waran text file a preliminary reason for encounter was identified in 883 of the 1035 lines (85.3 %). 1072 occurrences of 502 different reasons for encounter were found. 411 of the reasons for encounter were only found once (38% of occurrences, 82% of types) and 264 consist of one or two words.

The most common reasons for encounter in surgery, in both the total material and the Waran text file is buksmärtor (abdominal pains). The most common reasons for encounter in each file is shown in tables 1-2. Among patients of internal medicine, however, stated reasons for encounter are quite different. The most common all over is chest pain, and bleeding related disorders are rare, reflecting the different patient panorama in internal medicine – thus it can be expected that bleeding complications are presented at the dept of surgery even

Table 3 - Reasons for encounter in internal medicine

Medicine - total			
Reason for encounter		n	%
bröstsmärta	chest pain	6748	11,90
yrsel	dizziness	3311	5,84
huvudvärk	headache	2061	3,63
andnöd	dyspnea	2008	3,54
bröstsmärtor	chest pains	1951	3,44
dyspne	dyspnea	1851	3,26
brsm	chest pain	1398	2,47
feber	fever	1394	2,46
hjärtklappning	palpitations	1271	2,24
andfåddhet	dyspnea	1096	1,93
hosta	cough	1037	1,83
illamående	nausea	1022	1,80
intox	intoxication	905	1,60
trötthet	tiredness	779	1,37
förmaksflimmer	atrial fibrillation	730	1,29
andningsbesvär	breathing problems	696	1,23
syncope	syncope	640	1,13
kräkningar	vomiting	622	1,10
dvt	DVT	529	0,93
nedsatt at	poor general status	462	0,81

though the warfarin medication is given for diseases of neurology / cardiology at the dept of internal medicine.

Table 4 - Reasons for encounter in internal medicine,among Warfarin patients

Medicine - Waran		**occurrence**	
reason for encounter		n	%
bröstsmärta	chest pain	775	10,43
förmaksflimmer	atrial fibrillation	452	6,08
dyspne	dyspnea	435	5,85
andnöd	dyspnea	364	4,90
andfåddhet	dyspnea	327	4,40
yrsel	dizziness	326	4,39
hjärtklappning	palpitations	241	3,24
bröstsmärtor	chest pain	225	3,03
trötthet	tiredness	183	2,46
brsm	chest pain	167	2,25
dvt	DVT	146	1,96
feber	fever	138	1,86
illamående	nausea	106	1,43
hosta	cough	98	1,32
andningsbesvär	breathing difficulties	92	1,24
huvudvärk	headache	90	1,21
snabbt förmaks-flimmer	rapid atrial fibrillation	87	1,17
stroke	stroke	83	1,12
hjärtsvikt	heart failure	78	1,05
syncope	syncope	77	1,04

Table 5 - Ratio of frequencies of warfarin related surgical reasons for encounter (warfarin cases compared to total material).of individual reasons for encounter

Surgery - ratio		Frequency (‰)		Ratio W/
Sv	En	Total	Waran	Total
hematuri	hematuria	2,73	13,06	4,78
gastrointestinal blödning	GI bleeding	2,66	12,13	4,57
svart avföring	melena	2,20	9,33	4,25
rektalblödning	rectal bleeding	3,35	11,19	3,34
blödning per rektum	rectal bleeding	2,10	6,53	3,12
melena	melena	8,16	25,19	3,08
subileus	subileus	1,02	2,80	2,75
rektal blödning	rectal bleeding	2,38	6,53	2,75
makroskopisk hematuri	macroscopic hematuria	1,04	2,80	2,70
nedsatt at	general malaise	1,80	4,66	2,60
ikterus	icterus	2,91	7,46	2,56
Trötthet	tiredness	2,69	6,53	2,42
Anemi	anemia	3,25	7,46	2,29
hematemes	hematemesis	8,03	17,72	2,21
nedsatt allmäntillstånd	general malaise	1,28	2,80	2,19

urinstämma	urinary obstruction	4,17	8,40	2,01
Blodiga kräkningar	hematemesis	1,46	2,80	1,92
falltrauma	fall	12,16	23,32	1,92
urinretention	urinary retention	2,04	3,73	1,83
smärtor under höger arcus	pain upper right abdomen	1,04	1,87	1,80

Table 6 - Frequency ratios (warfarin/total) among internal medicine patients

Medicine - ratio		number total	number warfarin	Ratio W/ Total
nyupptäckt förmaksflimmer	new AF	15	15	7,5
sviktsymtom	heart failure	5	5	7,5
fladder	atrial flutter	4	4	7,5
högt inr	high INR	3	3	7,5
högt pk-värde	high PK	3	3	7,5
nydebuterad förmaksflimmer	new AF	3	3	7,5
recidiverande förmaksflimmer	recurring AF	3	3	7,5
smärta i vä skulderblad	shoulder pain	3	3	7,5
smärtor vänster underben	Pain left calf	3	3	7,5
susp cerebral insult	Suspected stroke	3	3	7,5
waranblödning	warfarin bleeding	3	3	7,5

To assess which reasons for encounter were more frequent among warfarin patients, we calculated the ratio of frequencies between the two patient groupsgroups, and the highest ratios are presented in table 2, in falling ratio – larger ratio means this problem is more common among warfarin treated patients. It is quite clear in the surgical material that bleeding related reasons for encounter are much more common among the warfarin patients than in the whole patient. In internal medicine, however, this is not so clear – there are many overfrequent problems but here the more frequent problems are rather related to indications for anticoagulation, such as atrial fibrillation and stroke, and no obvious relation to adverse effects can be seen.

Discussion

We have developed methods to identify and extract reasons for encounter from plain text in patient records of a department of general surgery at a teaching hospital. We find a reason for encounter in about 85 % of the records, suggesting that the method could be useful for example in statistical calculations and studies of patient groups. As a separate group we considered patients where "Waran" is mentioned. This is the trade name of warfarin in Sweden, the only oral anticoagulant currently on the market. Mention in the text does not necessarily imply that the patient is treated with warfarin before hospital admission, and this is a weakness in the design, but at least it suggests that patient has some condition where warfarin is discussed. Methods to ascertain whether current treatment includes warfarin would be very interesting, but this adds a lot of complexity to the information retrieval strategy. Experience from other authors suggest that general negation-detecting methods {{c:\fo\refs\ir.ref #120; Chapman et al., 2005}} can be used to distinguish "Waran" from "not Waran" which seems hopeful.

The possibilities of text data mining is getting a lot of recent attention within biomedicine, but most of the work is done in the domain of identifying genes and biochemical entities in published literature, much less work is focused on patient data[4]. Also coding of "chief complaint" has been attempted[5,6,7] . In Sweden, we also need to adapt all methods to our language, and we are not aware of any similar work done in Swedish.

Many of the reasons for encounter identified with this method are only found once and consist of long text strings. These reasons for encounter often result from a string matching where the rest of the sentence after words like inkommer för (comes because of?) is saved. Some of these lines and the lines where no reason for encounter was found need to be further analyzed and perhaps handled in a different way to make it possible to decide which the actual reasons for encounter are.

The result of the study could probably be improved by adding more patterns to the string matching scripts and by improving the already existing patterns. Analyzing the lines where a reason for encounter have not been found might help in doing this. A thorough analysis of the identified reasons for encounter is also needed to see what kinds of mistakes are made during identification, and what can be done to improve the precision and remove incorrect reasons. A larger text material would probably make the

4 Pakhomov, S.V., A. Ruggieri &C.G. Chute; Maximum entropy modeling for mining patient medication status from free text. *Proceedings / AMIA Symposium. Annual Symposium*, 587-591, 2002.

5 Mikosz, C.A., J. Silva, S. Black, G. Gibbs &I. Cardenas; Comparison of two major emergency department-based free-text chief-complaint coding systems. *MMWR. Morbidity and Mortality Weekly Report*, 53 Suppl, 101-105, 2004.

6 Shapiro, A.R; Taming variability in free text: application to health surveillance. *MMWR. Morbidity and Mortality Weekly Report*, 53 Suppl, 95-100, 2004.

7 Chapman, W.W., L.M. Christensen, M.M. Wagner, P.J. Haug, O. Ivanov, J.N. Dowling &R.T. Olszewski; Classifying free-text triage chief complaints into syndromic categories with natural language processing. *Artificial Intelligence in Medicine*, 33, 31-40, 2005.

numbers and calculations more accurate and would also be a way of further improving the results.

Many of the reasons for encounter are different morphological forms of the same word and should be considered synonyms, like buksmärtor and buksmärta (abdominal pain/pains). Natural language processing techniques, like part-of-speech tagging and stemming, can be applied to the material to help joining morphological forms. Work to accomplish this improvement has begun.

Grouping and categorizing the reasons for encounter could be achieved by looking for certain indicators of for example location, laterality or time. These indicators can then be used to search for reasons for encounter that are synonymous but not variants of the same word, like "hematemes" and "blodiga kräkningar" which refer to the same health condition. To accomplish this we hope to use medical morphemes from the Morphosaurus project[8], where we have been participating in entering Swedish morphemes.

It would be interesting to apply this method to different kinds of medical text and some work on this has already been started. The initial tests show that some modifications of the method are needed, and the pattern matching algorithm might have to be changed according to the text at hand.

Acknowledgments

Funding for the project has been provided by the European Commission within the SemanticMining Network of Excellence, the Center for Health Systems Analysis (CHSA), and from the Västra Götaland regional research fund.

Address for correspondence

Anders Thurin, Clinical Physiology, SU/Östra, SE-41685 GÖTEBORG, Sweden, anders.thuri@vgregion.se

8 Markó K, Schulz S, Hahn U: MorphoSaurus - Design and Evaluation of an Interlingua-based, Cross-language Document Retrieval Engine for the Medical Domain. *Method Inform Med*, 2005; 44 (4) : 537-545