Medical Informatics Europe 2002
G Surján et al. (eds)
IOS Press 2002

Shallow linguistic analysis of a large corpus of drug prescriptions

Anders THURIN[1], Mats WENNBERG[2], Karolina ANTONOV[2], Gunnar KLEIN[3]

[1]*Clinical Physiology, Sahlgrenska University Hospital,SE-416 85 Göteborg, Sweden,*
[2]*Apoteket AB, Stockholm, Sweden,* [3]*Karolinska Institute, Stockholm, Sweden*

Abstract. We report on first experiences from linguistic analyses of patient instructions from 198404 actual drug prescriptions regarding seven pharmaceutical products frequently prescribed in Sweden. The analysis includes expressions for amount, dose unit, dose interval, mode of administration, purpose and a few further details. Even simple processing seems useful to extract information from these short, rather formal text strings. We estimate the potential for calculation of prescribed dose from this material, and collected material gives a good starting point for more advanced linguistic analyses.

1. Introduction

An ongoing Swedish project aims to define a reference terminology for patient instructions in prescriptions. These kind of texts represent a simple stereotyped language, where relevant patterns can be found by statistical methods. Also numeric amounts are prevalent and important. There are several motives for doing such work: computer processable patient instructions can enable decision support and quality assurance, improve precision in pharmaceutical statistics, and fulfil legal requirements that patient instructions should be expressed in a way and in a language suitable for the patient (where an increasing number of people in Sweden don't have Swedish as their first language). The project is a collaboration between SIS – the Swedish Standards Institute and Apoteket AB – the national company running pharmacies.

2. Materials and Methods

The source material consisted of patient instructions as printed on a label on the drug when sold in a pharmacy in Sweden (all pharmacies are run by a government-owned company, Apoteket AB). The products analysed were:

Drug (Swedish trade name)	Drug form	Substance; Purpose	Number
Alvedon	Tabl 500mg	Paracetamol	46887
Bricanyl Turbohaler	Inhalation powder, 0,25mg/dose	Terbutaline inhalation powder	13592
Fenuril	Cream	Carbamide cream for dry skin	14595
Kåvepenin	Tabl 1g	Penicillin V	54929
Mollipect	Mixture	Cough syrup	37462
Nitromex	Resoribletter, 0,5mg	Sublingual nitroglycerin	14462
Xalatan	Eye drops, 50 µg/ml	Latanoprost; Eye drops for glaucoma	16477
			198404

The patient instructions are collected in a database and we analysed extracts from this database on all prescriptions given in a month in all of Sweden for these products.

In total, 198 404 prescriptions were analysed, and in the total material 50 423 different phrases occurred. The number of prescriptions for each drug form is shown in table 1. The drug form and strength was identified in other fields in the database and not examined here.

Analysis was done via a set of about 40 small custom programs written in Perl (Practical Extraction and Report Language), a language well suited for text analysis and with tools available for free [1]. Programs typically searched for patterns specified as regular expressions and extracted and/or counted occurrences of specified constructions. In addition, hypotheses were generated and tested by concordance analysis using Concordance 3.0 [2]. Simple statistics were performed in a spreadsheet program.

3. Results

3.1. Technical details

Most programs would run in a few seconds (using Windows, an Intel Pentium processor at 800-1400 MHz and 192-256MB memory). One concordance produced a file of 150MB, taking a few minutes to produce, about 15 s to load, but other operations in the program were quick.

3.2. Substitutions

Patient instructions were entered by pharmacists through a system which includes preprocessing of acronyms generating parts of phrases, such as TAR→TABLETTER (tablets) or GD→GÅNGER DAGLIGEN (times daily). 91 such phrases are available in the pharmacy system. Analysing the frequency of these phrases gives a hint on the use of the substitution function, but the phrases may also have been entered explicitly. We found around 408000 such phrases - 2,1 per patient instruction.

3.3. Word Frequency

In all, 5470 different words occur, 31 of these had >10000 occurrences, 97 >1000, 259 >100 and 838 >10 occurrences. 2998 words appear only once. The most common were:

Word	frequency	VID (at/for)	65022	10	29506
GÅNGER (times)	126756	TABLETT (tablet)	53426	DAGAR (days)	28170
DAGLIGEN (daily)	125494	I (in)	48641	4	27203
MOT (against)	81021	BEHOV (need)	47829	VÄRK (ache)	24736
2	80760	ML	36500	HOSTA (cough)	21600
1	78790	3	35405	15	20367
TABLETTER (tablets)	66645	1-2	35131	OCH (and)	17974

3.4. Common strings

Identical patient instructions often occurred in several prescriptions. 19 phrases occurred >1000 times, and the most common were:

Drug	Phrase	n
kavepenin	1 TABLETT 2 GÅNGER DAGLIGEN (2 times daily)	4465
kavepenin	1 TABLETT 2 GÅNGER DAGLIGEN I 10 DAGAR (for 10 days)	3406
kavepenin	2 TABLETTER 2 GÅNGER DAGLIGEN	3062
kavepenin	2 TABLETTER 2 GÅNGER DAGLIGEN I 10 DAGAR	2575
mollipect	15 ML 3 GÅNGER DAGLIGEN	1974
mollipect	15 ML 3 GÅNGER DAGLIGEN MOT HOSTA (against cough)	1923
mollipect	15 ML 3-4 GÅNGER DAGLIGEN MOT HOSTA	1600

kavepenin	1 TABLETT 2 GÅNGER DAGLIGEN MOT INFEKTION (against infection)	1563
kavepenin	2 TABLETTER 2 GÅNGER DAGLIGEN MOT INFEKTION	1263
kavepenin	1 TABLETT 3 GÅNGER DAGLIGEN	1211

3.5. Amount per dose

The initial word of the patient instruction was a numeric amount (expressed with words or numbers) in a large majority of cases - 89%.

3.6. Dose unit

This was often the second word of the patient instruction, after an initial numeric amount. It is expressed as "tablet" or a derivation thereof in a large majority of such drugs - for Alvedon (n=46887) in 45336 cases (97%), for Kåvepenin (n=54929) in 53654 cases (98%). Nitroglycerin is called tablet in 89% of cases, but can also called "resoriblett" in 4%. The inhalation powder Bricanyl (n=13592) is referred to in several ways:

Inhalationer (inhalations)	5096	37,49%	Puffar	248	1,82%
Inhalation	3804	27,99%	Inandning (inspiration)	125	0,92%
Doser (doses)	2135	15,71%	Inandningar	124	0,91%
Dos	1172	8,62%	Puff	74	0,54%

For the cough syrup Mollipect (n=37462) dose is expressed in ml (milliliters), in 88% of cases. Eye drops Xalatan (n=16477) are referred to as drop or drops in 90% of cases. For the crème Fenuril (n=14595) dose is rarely given, but thin application (TUNT) is mentioned 433 times, generous (RIKLIGT) 39 times.

3.7. Doses per day

The time pattern for dosage can be expressed in different ways: most common is a number of times per day, and the combination GÅNGER DAGLIGEN (times daily) is very common – about 125000 occurrences, PER DYGN (per 24 hours) occurs 7860 times, PER DAG (per day) 1628 times. Other ways to express time patterns are MORGON OCH KVÄLL (morning and evening) with around 11600 occurrences. Other common patterns are TILL NATTEN, TILL KVÄLLEN or PÅ KVÄLLEN (all meaning "at night") with 6337, 4489 and 2999 occurrences, respectively.

Dosage interval can also be given as a period of time, as in EVERY x HOURS, but this is less common with approx. 800 occurrences. Even less common are patterns involving fixed times – KL or KLOCKAN (Swedish markers for "at a point in time") occurring less than 300 times among the 198000 patient instructions.

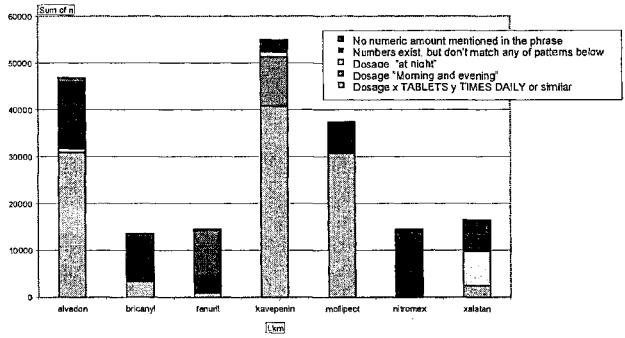

Figure 1

3.8. Calculation of dose

A few attempts were done via some simple Perl programs to extract the actual dosage prescribed from the patient instructions. This was successful in about 70% of cases, but this figure is very sensitive to method used and differences between drugs are large – see figure:

3.9. Maximal dose

The maximal number of doses per day are given in rather few cases, around 10000, as marked by the expressions HÖGST, MAX, UPP TILL and MAXIMALT (all denoting "maximally") with 6626, 2735, 922 and 251 occurrences, respectively.

3.10. Route

For tablets this is very rarely mentioned – the expressions PERORALT or GENOM MUNNEN (orally) occur only 2+65 times, most often regarding inhalation powder.
That Nitromex should be taken UNDER TUNGAN (sublingually) is mentioned i 8803 cases. For eyedrops Xalatan it is often mentioned which eye it should be used. VARDERA(each) occurs 7220 times, HÖGER(right) 2912 times, VÄNSTER(left) 2878 times and BÅDA(both) 1051 times.

Fenuril crème is applied in many ways SMÖRJES 616 times, PÅSTRYKES 575 times, APPLICERAS 534 times, INGNIDES 112 times, PÅSMÖRJES 108 times, INMASSERAS 47times, STRYKES 15 times, UTSTRYKES 13 times, ANBRINGAS 2times, APPLICERAS 2times.

3.11. Treatment duration

Is given mostly for the antibiotic Kåvepenin, in about half of the 53000 instances. 10 days is the most common duration, occuring in about 23000 cases.

3.12. Purpose of/reason for treatment

This is often expressed with a prepositional phrase:

MOT (against) occurs 81021 times, VID (at) 65022 times, but 47700 of these are VID BEHOV(when needed). FÖR (for) occurs 5752 times.

Common reasons according to this pattern (those occurring >1000 times), are:

Drug	Preposition	Reason	n
kavepenin	MOT	INFEKTION (infection)	9812
kavepenin	MOT	HALSFLUSS (pharyngitis)	1809
kavepenin	MOT	TANDINFEKTION (tooth infection)	1409
kavepenin	MOT	HALSINFEKTION (throat infection)	1043
kavepenin	VID	TANDINFEKTION	101
alvedon	MOT	VÄRK (chronic pain)	15871
alvedon	MOT	SMÄRTA (pain)	3341
alvedon	VID	VÄRK	3616
mollipect	MOT	HOSTA (cough)	13343
mollipect	VID	HOSTA	3060
bricanyl	MOT	ASTMA	2090
nitromex	MOT	KÄRLKRAMP	4323
nitromex	VID	KÄRLKRAMP	1896
fenuril	MOT	TORR HUD	2607

Another way of expressing the purpose of treatment is by verbs in participle form – this occurs 19290 times, and the most common are:

Reason	alvedon	bricanyl	fenuril	mollipect	Total
SLEMLÖSANDE (mucus clearing)				6637	6637
MJUKGÖRANDE (softening)			4399		4399
LUFTRÖRSVIDGANDE (bronchodilating)		1546		1150	2696
SMÄRTSTILLANDE (pain relieving)	1670				1670

4. Discussion

This rather simple method seems useful for analysis of this stereotypic text material, a lot of information can be extracted. Interesting previous studies have been done on drug indications mentioned in a drug dictionary [3], the current work is based on instructions from physician to individual patient, and has more focus on dosage. Some parts of phrases exhibit very small variation, probably in part due to the computer supported entry of instructions available at the pharmacies. For phrases not included in this system, misspellings are common and variation in phrasing instructions larger. Dose can often be calculated from these phrases, but is sometimes only implied or not mentioned. The phrases reflect some peculiarities of the Swedish language and Swedish medical jargon.

5. Conclusion

In spite of rather simple methodology in this analysis we can extract a fair amount of interesting data from a large sample of real-life patient instructions. Also a few cases of serious mistakes in patient instructions have been identified, and the results give a solid basis for describing a set of recommendations for patient instructions for future use e.g. in computer support for drug prescription.

References

[1] http://www.activestate.com, http://www.perl.com, http://www.perl.org
[2] http://www.rjcw.freeserve.co.uk/
[3] Duclos C, Venot A. Structured representation of drug indications: Lexical and semantic analysis and object-oriented modelling. Methods Inf Med 2000;**39**:83-7.
[4] Evans DA, Brownlow ND, Hersh WR, Campbell EM. Automating concept identification in the electronic medical record: an experiment in extracting dosage information. Proc AMIA Annu Fall Symp 1996;: 388-92