

# Exploiting Structured Data, Negation Detection and SNOMED CT Terms in a Random Indexing Approach to Clinical Coding

**Aron Henriksson**

DSV, Stockholm University  
aronhen@dsv.su.se

**Martin Hassel**

DSV, Stockholm University  
xmartin@dsv.su.se

## Abstract

The problem of providing effective computer support for clinical coding has been the target of many research efforts. A recently introduced approach, based on statistical data on co-occurrences of words in clinical notes and assigned diagnosis codes, is here developed further and improved upon. The ability of the word space model to detect and appropriately handle the function of negations is demonstrated to be important in accurately correlating words with diagnosis codes, although the data on which the model is trained needs to be sufficiently large. Moreover, weighting can be performed in various ways, for instance by giving additional weight to ‘clinically significant’ words or by filtering code candidates based on structured patient records data. The results demonstrate the usefulness of both weighting techniques, particularly the latter, yielding 27% exact matches for a general model (across clinic types); 43% and 82% for two domain-specific models (ear-nose-throat and rheumatology clinics).

## 1 Introduction

Clinicians spend much valuable time and effort in front of a computer, assigning diagnosis codes during or after a patient encounter. Tools that facilitate this task would allow costs to be reduced or clinicians to spend more of their time tending to patients, effectively improving the quality of healthcare. The idea, then, is that clinicians should be able simply to verify automatically assigned codes or to select appropriate codes from a list of recommendations.

## 1.1 Previous Work

There have been numerous attempts to provide clinical coding support, even if such tools are yet to be widely used in clinical practice (Stanfill et al., 2010). The most common approach has been to view it essentially as a text classification problem. The assumption is that there is some overlap between clinical notes and the content of assigned diagnosis codes, making it possible to predict possible diagnosis codes for ‘uncoded’ documents. For instance, in the *2007 Computational Challenge* (Pestian et al., 2007), free-text radiology reports were to be assigned one or two labels from a set of 45 ICD-9-CM codes. Most of the best-performing systems were rule-based, achieving micro-averaged  $F_1$ -scores of up to 89.1%.

Some have tried to enhance their NLP-based systems by exploiting the structured data available in patient records. Pakhomov et al. (2006) use gender information—as well as frequency data—to filter out improbable classifications. The motivation is that gender has a high predictive value, particularly as some categories make explicit gender distinctions.

Medical terms also have a high predictive value when it comes to classification of clinical notes (see e.g. Jarman and Berndt, 2010). In an attempt to assign ICD-9 codes to discharge summaries, the results improved when extra weight was given to words, phrases and structures that provided the most diagnostic evidence (Larkey and Croft, 1995).

Given the inherent practice of ruling out possible diseases, symptoms and findings, it seems important to handle negations in clinical text. In one study, it was shown that around 9% of automatically detected SNOMED CT *findings* and *disorders* were negated (Skeppstedt et al., 2011). In the attempt of Larkey and Croft (1995), negated medical terms are annotated and handled in various

ways; however, none yielded improved results.

## 1.2 Random Indexing of Patient Records

In more recent studies, the *word space model*, in its *Random Indexing* mold (Sahlgren, 2001; Sahlgren, 2006), has been investigated as a possible alternative solution to clinical coding support (Henriksson et al., 2011; Henriksson and Hassel, 2011). Statistical data on co-occurrences of words and ICD-10<sup>1</sup> codes is used to build predictive models that can generate recommendations for uncoded documents. In a list of ten recommended codes, general models—trained and evaluated on all clinic types—achieve up to 23% exact matches and 60% partial matches, while domain-specific models—trained and evaluated on a particular type of clinic—achieve up to 59% exact matches and 93% partial matches.

A potential limitation of the above models is that they fail to capture the function of negations, which means that negated terms in the clinical notes will be positively correlated with the assigned diagnosis codes. In the context of information retrieval, Widdows (2003) describes a way to remove unwanted meanings from queries in vector models, using a vector negation operator that not only removes unwanted strings but also synonyms and neighbors of the negated terms. To our knowledge, however, the ability of the word space model to handle negations has not been studied extensively.

## 1.3 Aim

The aim of this paper, then, is to develop the Random Indexing approach to clinical coding support by exploring three potential improvements:

1. Giving extra weight to words used in a list of SNOMED CT terms.
2. Exploiting structured data in patient records to calculate the likelihood of code candidates.
3. Incorporating the use of negation detection.

## 2 Method

Random Indexing is applied on patient records to calculate co-occurrences of tokens (words and

---

<sup>1</sup>The 10th revision of the *International Classification of Diseases and Related Health Problems* (World Health Organization, 2011).

ICD-10 codes) on a document level. The resulting models contain information about the ‘semantic similarity’ of individual words and diagnosis codes<sup>2</sup>, which is subsequently used to classify uncoded documents.

### 2.1 Stockholm EPR Corpus

The models are trained and evaluated on a Swedish corpus of approximately 270,000 clinically coded patient records, comprising 5.5 million notes from 838 clinical units. This is a subset of the *Stockholm EPR* corpus (Dalianis et al., 2009). A *document* contains all free-text entries concerning a single patient made on consecutive days at a single clinical unit. The documents in the partitions of the data sets on which the models are trained (90%) also include one or more associated ICD-10 codes (on average 1.7 and at most 47). In the testing partitions (10%), the associated codes are retained separately for evaluation. In addition to the complete data set, two subsets are created, in which there are documents exclusively from a particular type of clinic: one for *ear-nose-throat* clinics and one for *rheumatology* clinics.

Variants of the three data sets are created, in which negated clinical entities are automatically annotated using the Swedish version of *NegEx* (Skeppstedt, 2011). The clinical entities are detected through exact string matching against a list of 112,847 *SNOMED CT* terms belonging to the semantic categories ‘finding’ and ‘disorder’. It is important to handle ambiguous terms in order to reduce the number of false positives; therefore, the list does not include findings which are equivalent to a common non-clinical unigram or bigram (see Skeppstedt et al., 2011). A negated term is marked in such a way that it will be treated as a single word, although with its proper negated denotation. Multi-word terms are concatenated into unigrams.

The data is finally pre-processed: lemmatization is performed using the *Granska Tagger* (Knutsson et al., 2003), while punctuation, digits and stop words are removed.

### 2.2 Word Space Models

Random Indexing is performed on the training partitions of the described data sets, resulting in

---

<sup>2</sup>According to the *distributional hypothesis*, words that appear in similar contexts tend to have similar properties. If two words repeatedly co-occur, we can assume that they in some way refer to similar concepts (Harris, 1954). Diagnosis codes are here treated as words.

a total of six models (Table 1): two variants of the general model and two variants of the two domain-specific models<sup>3</sup>.

Table 1: The six models.

w/o negations	w/ negations
<i>General_Model</i>	<i>General_NegEx_Model</i>
<i>ENT_Model</i>	<i>ENT_NegEx_Model</i>
<i>Rheuma_Model</i>	<i>Rheuma_NegEx_Model</i>

### 2.3 Election of Diagnosis Codes

The models are then used to produce a ranked list of recommended diagnosis codes for each of the documents in the testing partitions of the corresponding data sets. This list is created by letting each of the words in a document ‘vote’ for a number of semantically similar codes, thus necessitating the subsequent merging of the individual lists. This ranking procedure can be carried out in a number of ways, some of which are explored in this paper. The starting point, however, is to use the semantic similarity of a word and a diagnosis code—as defined by the cosine similarity score—and the  $\text{idf}^4$  value of the word. This is regarded as our baseline model (Henriksson and Hassel, 2011), to which negation handling and additional weighting schemes are added.

### 2.4 Weighting Techniques

For each of the models, we apply two distinct weighting techniques. First, we assume a *technocratic* approach to the election of diagnosis codes. We do so by giving added weight to words which are ‘clinically significant’. That is here achieved by utilizing the same list of SNOMED CT findings and disorders that was used by the negation detection system. However, rather than trying to match the entire term—which would likely result in a fairly limited number of hits—we opted simply to give weight to the individual (non stop) words used in those terms. These words are first lemmatized, as the data on which the matching is performed has also been lemmatized. It will also allow hits independent of morphological variations.

We also perform weighting of the correlated ICD-10 codes by exploiting statistics generated

<sup>3</sup>ENT = Ear-Nose-Throat, Rheuma = Rheumatology.

<sup>4</sup>Inverse document frequency, denoting a word’s discriminatory value.

from the fixed fields of the patient records, namely *gender*, *age* and *clinical unit*. The idea is to use known information about a to-be-coded document in order to assign weights to code candidates according to plausibility, which in turn is based on past combinations of a particular code and each of the structured data entries. For instance, if the model generates a code that has very rarely been assigned to a patient of a particular sex or age group—and the document is from the record of such a patient—it seems sensible to give it less weight, effectively reducing the chances of that code being recommended. In order for an unseen combination not to be ruled out entirely, additive smoothing is performed. Gender and clinical unit can be used as defined, while age groups are created for each and every year up to the age of 10, after which ten-year intervals are used. This seems reasonable since age distinctions are more sensitive in younger years.

In order to make it possible for code candidates that are not present in any of the top-ten lists of the individual words to make it into the final top-ten list of a document, all codes associated with a word in the document are included in the final re-ranking phase. This way, codes that are more likely for a given patient are able to take the place of more improbable code candidates. For the general models, however, the initial word-based code lists are restricted to twenty, due to technical efficiency constraints.

### 2.5 Evaluation

The evaluation is carried out by comparing the model-generated recommendations with the clinically assigned codes in the data. This matching is done on all four possible levels of ICD-10 according to specificity (see Figure 1).



Figure 1: The structure of ICD-10 allows division into four levels.

## 3 Results

The general data set, on which *General\_Model* and *General\_NegEx\_Model* are trained and evaluated, comprises approximately 274,000 documents and 12,396 unique labels. The ear-nose-throat data

set, on which *ENT\_Model* and *ENT\_NegEx\_Model* are trained and evaluated, contains around 23,000 documents and 1,713 unique labels. The rheumatology data set, on which *Rheuma\_Model* and *Rheuma\_NegEx\_Model* are trained and evaluated, contains around 9,000 documents and 630 unique labels (Table 2).

data set	documents	codes
<i>General</i>	~274 k	12,396
<i>ENT</i>	~23 k	1,713
<i>Rheumatology</i>	~9 k	630

Table 2: Data set statistics.

The proportion of the detected clinical entities that are negated is 13.98% in the complete, general data set and slightly higher in the ENT (14.32%) and rheumatology data sets (16.98%) (Table 3).

### 3.1 General Models

The baseline for the general models finds 23% of the clinically assigned codes (exact matches), when the number of model-generated recommendations is confined to ten (Table 4). Meanwhile, matches on the less specific levels of ICD-10, i.e. partial matches, amount to 25%, 33% and 60% respectively (from specific to general).

The single application of one of the weighing techniques to the baseline model boosts performance somewhat, the fixed fields-based code filtering (26% exact matches) slightly more so than the technocratic word weighting (24% exact matches). The negation variant of the general model, *General\_NegEx\_Model*, performs somewhat better—up two percentage points (25% exact matches)—than the baseline model. The technocratic approach applied to this model does not yield any observable added value. The fixed fields filtering does, however, result in a further improvement on the three most specific levels (27% exact matches).

A combination of the two weighting schemes does not appear to bring much benefit to either of the general models, compared to solely performing fixed fields filtering.

### 3.2 Ear-Nose-Throat Models

The baseline for the ENT models finds 33% of the clinically assigned codes (exact matches) and 34% (L3), 41% (L2) and 62% (L1) at the less specific levels (Table 5).

Technocratic word weighing yields a modest improvement over the baseline model: one percentage point on each of the levels. Filtering code candidates based on fixed fields statistics, however, leads to a remarkable boost in results, from 33% to 43% exact matches. *ENT\_NegEx\_Model* performs slightly better than the baseline model, although only as little as a single percentage point (34% exact matches). Performance drops when the technocratic approach is applied to this model. The fixed fields filtering, on the other hand, similarly improves results for the negation variant of the ENT model; however, there is no apparent additional benefit in this case of negation handling. In fact, it somewhat hampers the improvement yielded by this weighting technique.

As with the general models, a combination of the two weighting techniques does not affect the results much for either of the ENT models.

### 3.3 Rheumatology Models

The baseline for the rheumatology models finds 61% of the clinically assigned codes (exact matches) and 61% (L3), 68% (L2) and 92% (L1) at the less specific levels (Table 6).

Compared to the above models, the technocratic approach is here much more successful, resulting in 72% exact matches. Filtering the code candidates based on fixed fields statistics leads to a further improvement of ten percentage points for exact matches (82%). *Rheuma\_NegEx\_Model* achieves only a modest improvement on L2. Moreover, this model does not benefit at all from the technocratic approach; neither is the fixed fields filtering quite as successful in this model (67% exact matches).

A combination of the two weighting schemes adds only a little to the two variants of the rheumatology model. Interesting to note is that the negation variant performs the same or even much worse than the one without any negation handling.

## 4 Discussion

The two weighting techniques and the incorporation of negation handling provide varying degrees of benefit—from small to important boosts in performance—depending to some extent on the model to which they are applied.

Model	Clinical Entities	Negations	Negations/Clinical Entities
<i>General_NegEx_Model</i>	634,371	88,679	13.98%
<i>ENT_NegEx_Model</i>	40,362	5,780	14.32%
<i>Rheuma_NegEx_Model</i>	20,649	3,506	16.98%

Table 3: *Negation Statistics*. The number of detected clinical entities, the number of negated clinical entities and the percentage of the detected clinical entities that are negated.

<b>Weighting</b>	<i>General_Model</i>				<i>General_NegEx_Model</i>			
	E	L3	L2	L1	E	L3	L2	L1
<i>Baseline</i>	0.23	0.25	0.33	0.60	0.25	0.27	0.35	0.62
Technocratic	0.24	0.26	0.34	0.61	0.25	0.27	0.35	0.62
Fixed Fields	0.26	0.28	0.36	0.61	0.27	0.29	0.37	0.63
Technocratic + Fixed Fields	0.26	0.28	0.36	0.62	0.27	0.29	0.37	0.63

Table 4: *General Models*, with and without negation handling. Recall (top 10), measured as the presence of the clinically assigned codes in a list of ten model-generated recommendations. E = exact match, L3→L1 = matches on the other levels, from specific to general. The baseline is for the model without negation handling only.

<b>Weighting</b>	<i>ENT_Model</i>				<i>ENT_NegEx_Model</i>			
	E	L3	L2	L1	E	L3	L2	L1
<i>Baseline</i>	0.33	0.34	0.41	0.62	0.34	0.35	0.42	0.62
Technocratic	0.34	0.35	0.42	0.63	0.33	0.33	0.41	0.61
Fixed Fields	0.43	0.43	0.48	0.64	0.42	0.43	0.48	0.63
Technocratic + Fixed Fields	0.42	0.42	0.47	0.64	0.42	0.42	0.47	0.62

Table 5: *Ear-Nose-Throat Models*, with and without negation handling. Recall (top 10), measured as the presence of the clinically assigned codes in a list of ten model-generated recommendations. E = exact match, L3→L1 = matches on the other levels, from specific to general. The baseline is for the model without negation handling only.

<b>Weighting</b>	<i>Rheuma_Model</i>				<i>Rheuma_NegEx_Model</i>			
	E	L3	L2	L1	E	L3	L2	L1
<i>Baseline</i>	0.61	0.61	0.68	0.92	0.61	0.61	0.70	0.92
Technocratic	0.72	0.72	0.77	0.94	0.60	0.60	0.70	0.91
Fixed Fields	0.82	0.82	0.85	0.95	0.67	0.67	0.75	0.91
Technocratic + Fixed Fields	0.82	0.83	0.86	0.95	0.68	0.68	0.76	0.92

Table 6: *Rheumatology Models*, with and without negation handling. Recall (top 10), measured as the presence of the clinically assigned codes in a list of ten model-generated recommendations. E = exact match, L3→L1 = matches on the other levels, from specific to general. The baseline is for the model without negation handling only.

## 4.1 Technocratic Approach

The technocratic approach, whereby clinically significant words are given extra weight, does result in some improvement when applied to all models that do not incorporate negation handling. The effect this weighting technique has on *Rheuma\_Model* is, however, markedly different from when it is applied to the other two corresponding models. It could potentially be the result of a more precise, technical language used in rheumatology documentation, where certain words are highly predictive of the diagnosis. However, the results produced by this model need to be examined with some caution, due to the relatively small size of the data set on which the model is based and evaluated.

Since this approach appears to have a positive impact on all of the models where negation handling is not performed, assigning even more weight to clinical terminology may yield additional benefits. This would, of course, have to be tested empirically and may differ from domain to domain.

## 4.2 Structured Data Filtering

The technique whereby code candidates are given weight according to their likelihood of being accurately assigned to a particular patient record—based on historical co-occurrence statistics of diagnosis codes and, respectively, age, gender and clinical unit—is successful across the board. To a large extent, this is probably due to a set of ICD-10 codes being frequently assigned in any particular clinical unit. In effect, it can partly be seen as a weighting scheme according to code frequency. There are also codes, however, that make gender and age distinctions. It is likewise well known that some diagnoses are more prevalent in certain age groups, while others are exclusive to a particular gender.

It is interesting to note the remarkable improvement observed for the two domains-specific models. Perhaps the aforementioned factor of frequently recurring code assignments is even stronger in these particular types of clinics. By contrast, there are no obvious gender-specific diagnoses in either of the two domains; however, in the rheumatology data, there are in fact 23 codes that have frequently been assigned to men but never to women. In such cases it is especially beneficial to exploit the structured data in patient

records. It could also be that the restriction to twenty code candidates for each of the individual words in the general models was not sufficiently large a number to allow more likely code candidates to make it into the final list of recommendations. That said, it seems somewhat unlikely that a code that is not closely associated with any of the words in a document should make it into the final list.

Even if the larger improvements observed for the domain-specific models may, again, in part be due to the smaller amounts of data compared with the general model, the results clearly indicate the general applicability and benefit of such a weighting scheme.

## 4.3 Negation Detection

The incorporation of automatic detection of negated clinical entities improves results for all models, although more so for the general model than the domain-specific models. This could possibly be ascribed to the problem of data sparsity. That is, in the smaller domain-specific models, there are fewer instances of each type of negated clinical entity (11.7 on average in ENT and 9.4 on average in rheumatology) than in the general model (31.6 on average). This is problematic since infrequent words, just as very frequent words, are commonly assumed to hold little or no information about semantics (Jurafsky and Martin, 2009). There simply is little statistical evidence for the rare words, which potentially makes the estimation of their similarity with other words uncertain. For instance, Karlgren and Sahlgren (2001) report that, in their TOEFL test experiments, they achieved the best results when they removed words that appeared in only one or two documents. While we cannot just remove infrequent codes, the precision of these suggestions are likely to be lower.

The prevalence of negated clinical entities—almost 14% in the entire data set—indicates the importance of treating them as such in an NLP-based approach to clinical coding. Due to the extremely low recall (0.13) of the simple method of detecting clinical entities through exact string matching (Skeppstedt et al., 2011), negation handling could potentially have a more marked impact on the models if more clinical entities were to be detected, as that would likely also entail more negated terms.

There are, of course, various ways in which one may choose to handle negations. An alternative could have been simply to ignore negated terms in the construction of the word space models, thereby not correlating negated terms with affirmed diagnosis codes. Even if doing so may make sense, the approach assumed here is arguably better since a negated clinical entity could have a positive correlation with a diagnosis code. That is, ruling out or disconfirming a particular diagnosis may be indicative of another diagnosis.

#### 4.4 Combinations of Techniques

When the technocratic weighting technique is applied to the variants of the models which include annotations of negated clinical entities, there is no positive effect. In fact, results drop somewhat when applied to the two domain-specific models. A possible explanation could perhaps be that clinically significant words that are constituents of negated clinical entities are not detected in the technocratic approach. The reason for this is that the application of the Swedish NegEx system, which is done prior to the construction and evaluation of the models, marks the negated clinical entities in such a way that those words will no longer be recognized by the technocratic word detector. Such words may, of course, be of importance even if they are negated. This could be worked around in various ways; one would be simply to give weight to all negated clinical entities.

Fixed fields filtering applied to the NegEx models has an impact that is more or less comparable to the same technique applied to the models without negation handling. This weighting technique is thus not obviously impeded by the annotations of negated clinical entities, with the exception of the rheumatology models, where an improvement is observed, yet not as substantial as when applied to *Rheuma\_Model*.

A combination of the technocratic word weighting and the fixed fields code filtering does not appear to provide any added value over the sole application of the latter weighting technique. Likewise, the same combination applied to the NegEx version does not improve on the results of the fixed fields filtering.

In this study, fine-tuning of weights has not been performed, neither internally or externally to each of the weighting techniques. It may, of course, be that, for instance, gender distinctions are more in-

formative than age distinctions—or vice versa—and thus need to be weighted accordingly. By the same token should the more successful weighting schemes probably take precedence over the less successful variants.

#### 4.5 Classification Problem

It should be pointed out that the model-generated recommendations are restricted to a set of properly formatted ICD-10 codes. Given the conditions under which real, clinically generated data is produced, there is bound to be some noise, not least in the form of inaccurately assigned and ill-formatted diagnosis codes. In fact, only 67.9% of the codes in the general data set are in this sense ‘valid’ (86.5% in the ENT data set and 66.9% in the rheumatology data set). As a result, a large portion of the assigned codes in the testing partition cannot be recommended by the models, possibly having a substantial negative influence on the evaluation scores. For instance, in the ear-nose-throat data, the five most frequent diagnosis codes are not present in the restricted result set. Not all of these are actually ‘invalid’ codes but rather action codes etc. that were not included in the list of acceptable code recommendations. A fairer evaluation of the models would be either to include such codes in the restricted result set or to base the restricted result set entirely on the codes in the data. Furthermore, there is a large number of unseen codes in the testing partitions, which also cannot be recommended by the models (358 in the general data set, 79 in the ENT data set and 39 in the rheumatology data set). This, on the other hand, reflects the real-life conditions of a classification system and so should not be eschewed; however, it is interesting to highlight when evaluating the successfulness of the models and the method at large.

### 5 Conclusion

The Random Indexing approach to clinical coding benefits from the incorporation of negation handling and various weighting schemes. While assigning additional weight to clinically significant words yields a fairly modest improvement, filtering code candidates based on structured patient records data leads to important boosts in performance for general and domain-specific models alike. Negation handling is also important, although the way in which it is here performed seems to require a large amount of training data

for marked benefits. Even if combining a number of weighting techniques does not necessarily give rise to additional improvements, tuning of the weighting factors may help to do so.

## Acknowledgments

We would like to thank all members of our research group, *IT for Health*, for their support and input. We would especially like to express our gratitude to Maria Skeppstedt for her important contribution to the negation handling aspects of this work, particularly in adapting her Swedish NegEx system to our specific needs. Thanks also to the reviewers for their helpful comments.

## References

- Hercules Dalianis, Martin Hassel and Sumithra Velupillai. 2009. The Stockholm EPR Corpus: Characteristics and Some Initial Findings. In Proceedings of ISHIMR 2009, pp. 243–249.
- Zellig S. Harris. 1954. Distributional structure. *Word*, 10, pp. 146–162.
- Aron Henriksson, Martin Hassel and Maria Kvist. 2011. Diagnosis Code Assignment Support Using Random Indexing of Patient Records — A Qualitative Feasibility Study. In Proceedings of AIME, 13th Conference on Artificial Intelligence in Medicine, pp. 348–352.
- Aron Henriksson and Martin Hassel. 2011. Election of Diagnosis Codes: Words as Responsible Citizens. In Proceedings of Louhi, 3rd International Workshop on Health Document Text Mining and Information Analysis.
- Jay Jarman and Donald J. Berndt. 2010. Throw the Bath Water Out, Keep the Baby: Keeping Medically-Relevant Terms for Text Mining. In Proceedings of AMIA, pp. 336–340.
- Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Pearson Education International, NJ, USA, p. 806.
- Jussi Karlgren and Magnus Sahlgren. 2001. From Words to Understanding. *Foundations of Real-World Intelligence*, pp. 294–308.
- Ola Knutsson, Johnny Bigert and Viggo Kann. 2003. A Robust Shallow Parser for Swedish. In Proceedings of Nodalida.
- Leah S. Larkey and W. Bruce Croft. 1995. Automatic Assignment of ICD9 Codes to Discharge Summaries. In PhD thesis University of Massachusetts at Amherst, Amherst, MA, USA.
- Serguei V.S. Pakhomov, James D. Buntrock and Christopher G. Chute. 2006. Automating the Assignment of Diagnosis Codes to Patient Encounters Using Example-based and Machine Learning Techniques. *J Am Med Inform Assoc*, 13, pp. 516–525.
- John P. Pestian, Christopher Brew, Pawel Matykiewicz, DJ Hovermale, Neil Johnson, K. Bretonnel Cohen and Wlodzislaw Duch. 2007. A Shared Task Involving Multit-label Classification of Clinical Free Text. In Proceedings of BioNLP 2007: Biological, translational, and clinical language processing, pp. 97–104.
- Magnus Sahlgren. 2001. Vector-Based Semantic Analysis: Representing Word Meanings Based on Random Labels. In Proceedings of Semantic Knowledge Acquisition and Categorization Workshop at ESS-LLI'01.
- Magnus Sahlgren. 2006. The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. In PhD thesis Stockholm University, Stockholm, Sweden.
- Maria Skeppstedt. 2011. Negation detection in Swedish clinical text: An adaption of Negex to Swedish. *Journal of Biomedical Semantics* 2, S3.
- Maria Skeppstedt, Hercules Dalianis and Gunnar H. Nilsson. 2011. Retrieving disorders and findings: Results using SNOMED CT and NegEx adapted for Swedish. In Proceedings of Louhi, 3rd International Workshop on Health Document Text Mining and Information Analysis.
- Mary H. Stanfill, Margaret Williams, Susan H. Fenton, Robert A. Jenders and William R Hersh. 2010. A systematic literature review of automated clinical coding and classification systems. *J Am Med Inform Assoc*, 17, pp. 646–651.
- Dominic Widdows. 2003. Orthogonal Negation in Vector Spaces for Modelling Word-Meanings and Document Retrieval. In Proceedings of ACL, pp. 136–143.
- World Health Organization. 2011. International Classification of Diseases (ICD). In *World Health Organization*. Retrieved June 19, 2011, from <http://www.who.int/classifications/icd/en/>.