Occupational Medicine 2020;**70:**203–206
Advance Access publication on 22 February 2020 doi:10.1093/occmed/kqaa022

Towards automated detection of psychosocial risk factors with text mining

L. Uronen[1], H. Moen[2], S. Teperi[3], K.-P. Martimo[4], J. Hartiala[5] and S. Salanterä[1]

[1]Department of Nursing Science, University of Turku, 20014 Turku, Finland, [2]Department of Future Technologies, University of Turku, 20014 Turku, Finland, [3]Department of Mathematics and Statistics, University of Turku, 20014 Turku, Finland, [4]Ilmarinen Mutual Pension Insurance Company, 00180 Helsinki, Finland, [5]Department of Medicine, University of Turku, 20014 Turku, Finland.

Correspondence to: L. Uronen, Department of Nursing Science, University of Turku, Joukahaisenkatu 3-5, 20520 Turku, Finland. Tel: +35 8505 319759; e-mail: leena.uronen@utu.fi

Background	Psychosocial risk factors influence early retirement and absence from work. Health checks by occupational health nurses (OHNs) may prevent deterioration of work ability. Health checks are documented electronically mostly as free text, and therefore the effect of psychological risk factors on working capacity is difficult to detect.
Aims	To evaluate the potential of text mining for automated early detection of psychosocial risk factors by examining health check free-text documentation, which may indicate medical statements recommending early retirement, prolonged sick leave or rehabilitation. Psychosocial risk factors were extracted from OHN documentation in a nationwide occupational health care registry.
Methods	Analysis of health check documentation and medical statements regarding pension, sick leave and rehabilitation. Annotations of 13 psychosocial factors based on the Prima-EF standard (PAS 1010) were used with a combination of unsupervised machine learning, a document search engine and manual filtering.
Results	Health check documentation was analysed for 7078 employees. In 83% of their health checks, psychosocial risk factors were mentioned. All of these occurred more frequently in the group that received medical statements for pension, rehabilitation or sick leave than the group that did not receive medical statement. Documentation of career development and work control indicated future loss of work ability.
Conclusions	This study showed that it was possible to detect risk factors for sick leave, rehabilitation and pension from free-text documentation of health checks. It is suggested to develop a text mining tool to automate the detection of psychosocial risk factors at an early stage.
Key words	Health check; occupational health; psychosocial risk factors; text mining.

Introduction

In 2016, about 7% of working-age people in Finland received a disability pension [1]. A medical statement by an occupational physician is required for a pension, prolonged sick leave or rehabilitation [2]. Approximately 10–15% of disability pensions are related to psychosocial factors such as workload, computer skills and low job control [3]. Hazardous conditions, job demands and shift work are related to long-term sickness absence [4]. The European Prima-EF consortium framework, publicly available as PAS 1010, is a psychosocial risk management guide. The management guide recognizes 13 psychosocial factors (Table 1) [5,6].

Occupational health services aspire to prevent disability using job-specific health checks [6] whereby an occupational health nurse (OHN) gathers information about work ability and health. The documentation from theses health checks is appointment-specific and fragmented and does not enable easy assessment of employees' work ability [7].

Our aim was to identify psychosocial risk factors in health check documents using text mining to automate the identification process. We were interested in risk factors including recommendations for pension, prolonged sick leave or rehabilitation. The PAS 1010 classification was used to identify psychosocial risk factors.

© The Author(s) 2020. Published by Oxford University Press on behalf of the Society of Occupational Medicine.
All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from https://academic.oup.com/occmed/article/70/3/203/5747763 by guest on 26 August 2021

Key learning points

What is already known about this subject:

- Psychosocial risk factors influence early retirement and absence from work.
- Physical workload, computer skills and low job control are already known as psychosocial risk factors.
- Occupational health care records are appointment-specific and fragmented, which made it difficult to evaluate work ability.

What this study adds:

- Text mining can identify psychosocial risk factors described in the free text of health check records.
- Psychosocial risk factors documented in occupational health checks showed significant correlations with medical statements for pensions, prolonged sick leave or rehabilitation.
- Workload and career development were the most frequently documented categories of psychosocial risk factors (from a total of 13 categories).

What impact this may have on practice or policy:

- Charting psychosocial risk factors in health checks is important when predicting the future ability to work.
- A text mining tool could be used to detect psychosocial risk factors in an early stage in health checks.
- Text mining and artificial intelligence can help to identify psychosocial risk factors in occupational health records and help to start earlier interventions to prevent work disability.

Table 1. Total mentions of psychosocial risk factors in health check documentation

Psychosocial risk factor	Example	Total mentions (*n*)
Job content	Cannot cope, customer work	58
Violence in the workplace	Injustice, insecurity	10
Bullying at work	Bullying	10
Career development	Co-negotiation, short work breaks, stressful	4659
Work environment and equipment	Noise, injury at work, asbestos	1616
Harassment at work		0
Organizational culture and function	No support, no information, no contradiction	3019
Role in organization	Trust, lack of confidence	8
Interpersonal relationships at work	Accusation, not appreciated, self-acting	1494
Work schedule	Shift work, long shifts, no holidays	2795
Work control	Inadequacy, work management, one-sidedness	1202
Home-work interface	Have little time, no time for hobbies, loneliness	1590
Workload and work pace	No time, a lot of overwork, obsession	10 663

Methods

The data set used in this retrospective study contained 15 443 health check documents from 7078 employees from 2002 to 2007 produced by a nationwide private occupational health service in Finland.

The psychosocial risks included in the PAS 1010 were first manually annotated in a subset of documents from 196 employees. Annotated words and phrases associated with each of the 13 risk factors were extracted as lists (see Table 1). These lists were then used to detect additional words and phrases in a semi-supervised fashion with the help of a prototype search tool.

The prototype search tool consists primarily of a word similarity (semantic) model. This model was trained using the Word2vec toolkit [8] on the full data set, using the Skip-gram architecture, hierarchical softmax, a dimensionality of 300, and otherwise default parameters. For each word in the initial lists of risk factors, the model was used to identify and suggest additional words with similar meanings. This is comparable to the interactive rapid vocabulary exploration used in other studies [9]. For multi-word phrases, we found that in most cases, all words except the last were likely to be negations or prepositions (e.g. 'no information'). A simple heuristic was applied that assumes phrases of similar meaning can be identified by replacing the last word with words of similar meanings. The Apache Solr search platform was used to verify that the extracted candidates exist in the data set.

The extended lists of risk factors were used to create a table of the frequency of each risk factor.

Downloaded from https://academic.oup.com/occmed/article/70/3/203/5747763 by guest on 26 August 2021

Table 2. Average number of risk factors per person per health check

Variable	Medical statement	*n*	Amount of health checks	Job content	Violence in the workplace	Bullying at work	Career development	Work environment and equipment	Organizational culture and function	Role in organization	Interpersonal relationships at work	Work schedule	Work control	Home-work interface	Workload and work pace
Pension	No	7010	2.492	0.003	0.000	0.000	0.246	0.092	0.250	0.000	0.034	0.161	0.047	0.077	0.458
	Yes	68	4.265	0.000	0.001	0.010	0.895	0.082	0.770	0.000	0.139	0.209	0.456	0.213	1.301
	P-value			N/A	N/A	N/A	***	0.8088	***	N/A	***	*	***	**	***
Rehabilitation	No	6196	2.349	0.002	0.000	0.000	0.240	0.091	0.219	0.000	0.030	0.161	0.043	0.068	0.404
	Yes	882	2.254	0.009	0.001	0.001	0.343	0.108	0.000	0.002	0.084	0.162	0.104	0.153	1.068
	P-value			**	0.6236	0.5686	***	***	***	N/A	***	**	***	***	***
Sick leave	No	6750	2.416	0.003	0.000	0.000	0.249	0.093	0.245	0.000	0.032	0.162	0.042	0.072	0.452
	Yes	328	2.335	0.004	0.000	0.003	0.282	0.073	0.439	0.001	0.084	0.126	0.230	0.221	0.784
	P-value			0.6860	N/A	N/A	**	0.4458	***	N/A	***	0.5614	***	***	***

*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

Statistical analysis was performed using SAS version 9.4 for Windows (SAS Institute Inc., Cary, NC, USA). The significance of associations of independent and dependent variables (see Table 2) was determined by logistic regression. Variables were reported as medians and frequencies. The statistical significance level was set at $P = 0.05$; all tests were two-tailed.

This study has a permission and ethical approval from the Ministry of Social Affairs and Health in Finland.

Results

A total of 7078 employees were included, and 1820 medical statements were evaluated comprising 883 men and 937 women. The average age at the first medical statement was 42.5 years, and the average time from the first health check to the medical statement was 2.3 years.

The analysis found a total of 26 309 risk factors. A total of 284 unique risk factor words or phrases were identified and allocated to the 13 PAS 1010 categories, as shown in Table 1. Most psychosocial risk factors were in the workload and work pace category (10 663 mentions). There were 1171 (17%) employees whose documents contained no mention of psychosocial risk factors.

Sixty-eight workers received medical statements for retirement. The most commonly documented risk factors were workload and work pace (mean = 1.301), career development (mean = 0.895) and organizational culture and function (mean = 0.770), all of which occurred more often in workers with medical statements for retirement than in those who did not have such statements ($P \leq 0.0001$). When all psychosocial risk factors were considered, career development and work control were clear risks for early retirement ($P \leq 0.0001$). Compared to health checks that were not followed by medical statements for disability pension, workload and work pace were found up to 17.4 times more frequently in the health checks proceeding medical statements for pension (Table 2).

Eight hundred and eighty-two (882) workers had medical statements for rehabilitation. The most frequently mentioned category of psychosocial risk factors was workload and work pace (mean = 1.068). The differences between workers with and without medical statements for rehabilitation were all significant ($P \leq 0.0001$) (Table 2).

A total of 328 workers received medical statements for prolonged sick leave (Table 2). The psychosocial risk factor categories proceeding medical statements about sick leave were workload and work pace (mean = 0.784) and organizational culture and function (mean = 0.439). The differences between workers with and without medical statements were all significant ($P \leq 0.0001$) (Table 2).

Downloaded from https://academic.oup.com/occmed/article/70/3/203/5747763 by guest on 26 August 2021

Discussion

A key finding was that text mining enabled identification of psychosocial risk factors from free-text documentation. Psychosocial risk factors appeared more frequently in health checks proceeding medical statements than in those that did not. It is possible to develop a text mining tool to detect risk factors at an early stage. The results supported the development of structured electronic documentation of psychosocial risk factors in health checks.

OHN documentation of psychosocial risk factors could predict the loss of work ability. Job control is a known risk factor for early retirement [7–9], but less is known about career development as a psychosocial risk. The data in this study represented most branches and ages and came from a nationwide private OH service provider. One weakness of this study was that the number of medical statements was small, especially for disability pension.

Competing interests

None declared.

References

1. Statistical Yearbook of Pensioners in Finland. 2016. https://www.etk.fi/wp-content/uploads/statistical-yearbook-of-pensioners-in-finland-2016.pdf (26 June 2019, date last assessed).
2. National Pensions Act 11.5.2007/568. http://www.finlex.fi/fi/laki/ajantasa/2007/20070568 (26 June 2019, date last assessed).
3. Christensen KB, Feveile H, Labriola M, Lund T. The impact of psychosocial work environment factors on the risk of disability pension in Denmark. *Eur J Public Health* 2008;**18**:235–237.
4. Slany C, Schütte S, Chastang J-F, Parent-Thirion A, Vermeylen G, Niedhammer I. Psychosocial work factor and long sickness absence in Europe. *Int J Occup Environ Health* 2014;**20**:16–25.
5. Leka S, Cox T. *Guide Prima-EF. Guidance on the European Framework for Psychosocial Risk Management. A Resource for Employers and Workers Representatives*. Helsinki: WHO, 2011. http://www.prima-ef.org/uploads/1/1/0/2/11022736/prima-ef_brochure_english.pdf (26 June 2019, date last assessed).
6. Occupational Health Care Act 1383/2011. https://www.finlex.fi/fi/laki/ajantasa/2001/20011383 (26 June 2019, date last assessed).
7. Uronen L, Heimonen J, Puukka P, Martimo KP, Hartiala J, Salanterä S. Health check documentation of psychosocial factors using the WAI. *Occup Med (Lond)* 2017;**67**:151–154.
8. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. *Adv Neural Inf Process Syst* 2013;**26**:3111–3119.
9. Topaz M, Murga L, Gaddis KM *et al*. Mining fall-related information in clinical notes: comparison of rule-based and novel word embedding-based machine learning approaches. *J Biomed Inform* 2019;**90**:103103.

Downloaded from https://academic.oup.com/occmed/article/70/3/203/5747763 by guest on 26 August 2021