Original article

Scand J Work Environ Health. 2020;46(2):168–176. doi:10.5271/sjweh.3834

Predicting future changes in the work ability of individuals receiving a work disability benefit: weighted analysis of longitudinal data

by Ilse Louwerse, MSc,[1, 2, 3] Maaike A Huysmans, PhD,[1, 3] Jolanda HJ van Rijssen, PhD,[1, 2, 3] Frederieke G Schaafsma, MD, PhD,[1, 3] Kristel HN Weerdesteijn, MD,[1, 2, 3] Allard J van der Beek, PhD,[1, 3] Johannes R Anema, MD, PhD [1, 3]

Louwerse I, Huysmans MA, van Rijssen JHJ, Schaafsma FG, Weerdesteijn KHN, van der Beek AJ, Anema JR. Predicting future changes in work ability of individuals receiving a work disability benefit: weighted analysis of longitudinal data. *Scand J Work Environ Health*. 2020;46(2):168–176. doi:10.5271/sjweh.3834

Objectives Weighted regression procedures can be an efficient solution for cohort studies that involve rare events or diseases, which can be difficult to predict, allowing for more accurate prediction of cases of interest. The aims of this study were to (i) predict changes in work ability at one year after approval of the work disability benefit and (ii) explore whether weighted regression procedures could improve the accuracy of predicting claimants with the highest probability of experiencing a relevant change in work ability.

Methods The study population consisted of 944 individuals who were granted a work disability benefit. Self-reported questionnaire data measured at baseline were linked with administrative data from Dutch Social Security Institute databases. Standard and weighted multinomial logit models were fitted to predict changes in the work ability score (WAS) at one-year follow-up. McNemar's test was used to assess the difference between these models.

Results A total of 208 (22%) claimants experienced an improvement in WAS. The standard multinomial logit model predicted a relevant improvement in WAS for only 9% of the claimants [positive predictive value (PPV) 62%]. The weighted model predicted significantly more cases, 14% (PPV 63%). Predictive variables were several physical and mental functioning factors, work status, wage loss, and WAS at baseline.

Conclusion This study showed that there are indications that weighted regression procedures can correctly identify more individuals who experience a relevant change in WAS compared to standard multinomial logit models. Our findings suggest that weighted analysis could be an effective method in epidemiology when predicting rare events or diseases.

Key terms prognosis; rare event; weighted multinomial logit model; work disability allowance.

Long-term work disability is bad for an individual's health, and returning to work is generally associated with a positive effect on the future course of the disease and work ability (1–3). Individuals who are unable to work due to a disease or disorder can apply for a work disability benefit. In most European countries, this covers both financial support to compensate loss of income and interventions supporting return to work.

Possible predictors for work disability include a broad range of external and personal factors. When conducting medical disability assessments to evaluate whether a work disability benefit should be granted, insurance physicians (IP) predominantly rely on factors relating to the disease and the disorder of a claimant (4, 5). One of the main tasks of an IP is estimating prognosis of work disability and determining if and when a reassessment should be planned (6). Medical reassessments are conducted to determine whether an individual's health has improved or deteriorated to such an extent that adjustment of support to return to work is required or the continuing eligibility for the benefit has changed. In The Netherlands, claim duration for work disability benefits is long lasting for

1 Amsterdam UMC, Vrije Universiteit Amsterdam, Department of Public and Occupational Health, Amsterdam Public Health Research Institute, Amsterdam, The Netherlands.
2 Dutch Institute of Employee Benefit Schemes (UWV), Amsterdam, The Netherlands.
3 Research Center for Insurance Medicine, AMC-UMCG-VUmc-UWV, Amsterdam, The Netherlands.

Correspondence to: Ilse Louwerse, Amsterdam UMC, Vrije Universiteit Amsterdam, Department of Public and Occupational Health, Amsterdam Public Health research institute, Van der Boechorststraat 7, NL-1081 BT Amsterdam, The Netherlands. [E-mail: i.louwerse@vumc.nl]

most claimants and IP consider prognosis of work disability as the most difficult part of the work disability assessment (7, 8). Therefore, accurate prognosis of future changes in work disability is important to identify those in need of return-to-work interventions and for efficient planning of medical reassessments.

Work ability, commonly measured with the Work Ability Index (WAI), is an important concept in the context of work disability duration. It is defined as the physical, mental, and social fit of an individual with the work demands and capability to participate in work (9). Self-assessed work ability is a strong predictor of work disability duration and return to work (10, 11). Clinical decision-support systems, in which characteristics of individual patients are used to generate patient-specific assessments or recommendations that are then presented to clinicians for consideration, are designed to aid decision-making (12). They can optimize the time with the client and improve the overall quality of services (13). A prediction model for future changes in work ability could aid IP in their medical disability assessment and lead to more precise estimation of future work disability. Because resources to perform medical reassessments are limited, the model is of most added value in practice if it can sufficiently identify claimants who will improve in their work ability. This ensures that medical reassessments are planned at the time an assessment interview with an IP has the most added value. However, claimants who perceive a relevant future improvement of their work ability form only a relatively small proportion of the total number of work disability claimants.

Predicting rare events or diseases with probabilistic statistical regression is difficult as these methods tend to be biased towards the majority class and underestimate the probability of rare events (14). Weighted regression can take account of the preponderance of claimants not experiencing a substantial change in their work ability, and focus accuracy on claimants who most likely will experience a change. Weighted least squares have its origin in econometrics and are used in a range of application areas, such as psychology, regional science and time series analysis (15, 16). However, we are not aware of any research in occupational epidemiology using weighted analysis. Therefore, the aim of this study was twofold: to (i) predict changes in work ability of claimants at one year after approval of the work disability benefit by building a model based on socio-demographic, work disability, health, functional limitation and personal factors; and (ii) explore whether the accuracy of predicting claimants with the highest probability of experiencing a relevant change of work ability could be improved by using weighted regression.

Methods

Study population

We used data of the FORWARD study, a longitudinal cohort study among 2539 individuals who applied for a work disability benefit at the Dutch Social Security Institute (SSI) between July 2014 and March 2015, after a two-year period of sick leave. Individuals were aged 18–64 years at inclusion. Claimants suffering from severe mental, cognitive, or visual disorders or those diagnosed with cancer were excluded from the FORWARD study. A more extensive description of the study cohort can be found elsewhere (Weerdesteijn et al. Does self-perceived health correlate with physician-assessed functional limitations in a medical work disability assessment? Submitted for publication, 2019).

From the FORWARD study, we retrieved data from the baseline questionnaire completed just before the medical disability assessment and the questionnaire at one-year follow-up. For each participant, we combined the self-reported data of the cohort study with administrative data from SSI databases. The participants of the FORWARD study all signed informed consent. The Medical Ethics Committee of the VU University Medical Center (Amsterdam, The Netherlands) has approved the FORWARD study.

Inclusion and exclusion criteria

To be included in the present study, the single-item question of the WAI needed to be answered both at baseline and one-year follow-up. Of the 2593 individuals included in the FORWARD study, 42 and 646 participants were excluded because they did not answer this question at baseline and one-year follow-up, respectively. We excluded participants who were ineligible or did not apply for work disability benefits (N=701) and those who were granted a permanent work disability benefit (N=260). In the latter case, there are no possibilities to return to work, and hence no reassessments need to be scheduled. In total, 944 participants were included in the present study.

Dependent variable

The dependent variable of the model was the change in self-reported work ability at one-year follow-up as compared to baseline. Work ability was measured with the first question of the WAI, also referred to as the work ability score (WAS) (17). This question asks participants to compare their current work ability with their lifetime best on a 0–10 scale. Higher scores indicate better work ability. The WAS is significantly correlated to the WAI and can therefore be used as a simple indicator for assess-

ing work ability (18, 19). A single-item measure takes less time to complete and analyze and is, therefore, preferable in terms of costs, interpretation and missing data.

In line with previous studies, we defined an improvement or deterioration in WAS of ≥2 points as the smallest detectable self-reported change likely to have an effect on job opportunities and work disability benefit (20, 21). Based on their change in WAS scores at one-year follow-up as compared to baseline, we divided the participants into three groups: participants with no relevant change ($|WAS_{T1} - WAS_{T0}| \leq 1$), an improvement ($WAS_{T1} - WAS_{T0} \geq 2$), or a deterioration ($WAS_{T1} - WAS_{T0} \leq -2$), with WAS_{T0} and WAS_{T1} the scores at baseline and one-year follow-up. WAS_{T0} was also added as an independent variable to the model.

Independent variables

All independent variables were measured at baseline. The socio-demographics age, gender, marital status, and educational level, as well as the work-related characteristics work status and occupational sector were retrieved from the SSI database. In addition, a number of health characteristics were determined: primary diagnosis, comorbidity, permanency, treatment and medication, and functional limitations as registered by the IP during the medical disability assessment in the list of functional abilities (LFA). The LFA is partly based on the World Health Organisation's International Classification (ICF) of Functioning, Disability, and Health (22). It consists of 106 items indicating the presence (dichotomous) and severity (ordinal) of limitations, categorized into six sections: personal functioning, social functioning, adjusting to the physical environment, dynamic movements, static posture, and working hours. Higher scores on the ordinal rating scales indicate more severe limitations to perform activities. We considered the average number of limitations of the first five sections and the single question of the last section regarding restrictions in the working hours per day as independent variables. If a claimant is too seriously disabled to return to work, eg, bedridden or receiving inpatient care, limitations are not registered in the LFA. This was the case for 119 (13%) of the participants in our study sample.

Besides registration data from the SSI, a number of self-reported surveys from the FORWARD study baseline questionnaire was used. The Short Form Health Survey (SF-36) is a measure of health status, containing 36 items on physical and mental functioning and role limitations, well-being, pain, general health, and health change. Scores range between 1–60, higher scores indicating better health status (23). The Whitely Index (WI) contains 14 items to measure health anxiety. Scores range between 0–14, with higher scores indicating more severe health anxiety (24). The Hospital Anxiety and Depression Scale (HADS) produces scales for anxiety and depression. Scores range between 0–21, with higher scores indicating higher distress (25). The Work and Well-being Inventory (WBI) measures symptoms, coping, support, stress, and disability with 87 items. Scores range between 0–84, with higher scores indicating more barriers for return to work (26). We also retrieved household and work-related characteristics, the latter regarding work demands and managerial tasks. The questionnaire also asked respondents about their expectations with respect to recovering and getting back to work.

Statistical analysis

Multinomial logistic regression analysis was used to predict changes in work disability at one-year follow-up. We fitted both standard and non-parametric multinomial logit (MNL) estimates. See figure 1 for the specification of the non-parametric MNL estimates. Because we were most interested in accurately predicting the largest improvements in WAS, we used the following linear weight function for claimants who experience an improvement in WAS (ie, $WAS_{T1} - WAS_{T0} \geq 2$):

$$w_i = \tfrac{1}{2}(WAS_{T1} - WAS_{T0}) + 1$$

For all claimants who did not experience an improvement in WAS (ie, $WAS_{T1} - WAS_{T0} < 2$), the weight was set to $w_i = 1$. By using the above linear weight function, claimants with an improvement in WAS of 2 points were assigned twice as much weight as claimants not experiencing an improvement in WAS. Because larger weights were assigned to claimants with a larger improvement in WAS, the model focusses on accurately predicting these claimants. In application areas where weighted regression is more often used, weight functions are often linear or exponential functions. For instance, in geographically weighted regression, locations that are closer get higher weights. In time series analysis, weights decrease for observations further back in time. Hence, the linear weight function of the present study is in line with weight specifications in other research (15, 27). Because weighted regression procedures are not commonly used in occupational epidemiology, there is no general approach to specify the exact weights that should be given to observations. Hence, we tried several weight functions and examined the effect on the performance of the prediction model. Assigning a weight equal to one to claimants with an improvement of WAS would result in the standard MNL model. Therefore, we considered assigning weights equal to 1.5, 2, 2.5, or 3. We did not consider weights >3 as we felt this would place disproportional emphasis on claimants with an improvement in WAS. Although the differences between the weight functions were small, we chose a weight of 2 in the final model as this resulted in

Let j = 1, 2, 3 denote the alternative categories that a claimant can belong to, based on the change in work ability at one-year follow-up, and let i = 1, … ,n denote the claimants. The probabilities p_{ij} of claimant i belonging to category j of the multinomial logit model are

$$p_{ij} = \text{Prob}[Y_i = j|x_i] = \frac{\exp(x_i'\beta_j)}{1 + \sum_{k=2}^{J}\exp(x_i'\beta_k)} \quad (1)$$

where x_i represents the characteristics of claimant i, and β_j measures the relative weights of the characteristics. The multinomial logit model can be estimated by maximum likelihood, ie, by maximizing the log-likelihood

$$\log(L) = \sum_{i=1}^{n}\sum_{j=1}^{3} I_{ij}\log(p_{ij}) \quad (2)$$

with respect to the parameters β_j, j = 1, 2, 3. Here, I_{ij} is an indicator variable, with I_{ij} = 1 if $Yi = j$ and I_{ij} = 0 otherwise.

Now, let w_i be the weight given to claimant i. Instead of minimizing (1) we could minimize the following pseudo log-likelihood function

$$\log(L) = \sum_{i=1}^{n} w_i\left[\sum_{j=1}^{3} I_{ij}\log(p_{ij})\right]. \quad (3)$$

Note that (3) includes standard multinomial logit as the special case with weights w_i = 1 for all values of i.

Figure 1. Specification of the non-parametric MNL estimates

the highest sensitivity, ie, the model that could identify most claimants with an improvement in WAS. The positive predictive value (PPV) and negative predictive value (NPV) were similar for the different weight functions that were considered.

The models were built using three steps. First, we performed univariable analyses to test the association of each independent variable with the outcome variable using likelihood ratio (LR) tests (cut off score P>0.2). Second, the variables remaining from the univariable analyses were tested for multicollinearity using variance inflation factors (VIF). We considered VIF <10 to be acceptable (28). Third, we selected the subset of predictors for the final model using a hybrid approach combining forward and backward selection procedures.

Before the start of the analysis, we randomly split the data into a training set (80% of the study population) to fit the models and a test set (20% of the study population) to evaluate the models. The purpose of developing the prediction model is that it can be used in practice. This means that we want to know how well the model predicts new cases. Therefore, the test set, ie, the held-out sample, is used to get an unbiased estimate of model effectiveness.

We calculated several performance measures to compare the standard and weighted MNL model. We reported both specificity and sensitivity as these are important measures of diagnostic accuracy of a model. However, they are of no practical use when IP need to estimate the probability of improvement in WAS for individual claimants (29). Hence, predictive values are more meaningful performance measures in this context. In general, there is a trade-off between sensitivity and predictive values. We can indicate the added value of the weighted model if it results in predictions with both higher sensitivity and predictive values.

We used McNemar's test to statistically assess whether the standard and the weighted model had a similar proportion of errors on the test set. Calculation of the test statistic is based on the contingency table. It tests whether the models have equal accuracy for predicting true improvements in WAS, ie, it detects whether the difference between the misclassification rates of the models is statistically significant. The level of significance was set at P<0.05.

All analyses were performed in RStudio for Windows, version 0.99.902.

Results

Tables 1 and 2 show the baseline characteristics of the study population. Mean WAS on baseline was 2.5 [standard deviation (SD) 2.1], and 2.8 (SD 2.2) at one-year follow-up. The majority of the study population (N=599; 63%) did not experience a change in WAS at one-year follow-up; 208 claimants (22%) experienced an improvement in WAS [mean WAS improvement 3.1 (SD 1.5)] and 127 a deterioration (15%).

In this section, we mainly focus on the results of the 187 claimants who were randomly selected to be included in the test set. Among this group, the percentage experiencing a WAS improvement at one-year follow-up was slightly higher than that of the training set (24% versus 21%). Of all cases in the test set, the standard model predicted for 16.9% of the total number of claimants, an improvement of the WAS at one-year follow-up (table 3). The sensitivity was only 22%,

Table 1. Descriptive statistics of the study population at baseline [CI= confidence intervals; LFA=list of functional abilities; MNL=multinomial logit; SD=standard deviation.]

	Study population (N=944)				Standard MNL model	Weighted MNL model
	N	%	Mean	SD	Coefficient (95% CI)	Coefficient (95% CI)
Occupational						
Work status (working)	200	21			1.03 (0.54–0.52)	1.03 (0.64–1.42)
Health						
Mental healthcare (yes)	487	52				-0.18 (-0.51–0.17)
Disability assessment						
Wage loss (≥ 80%)	548	58				-0.61 (-0.98–-0.25)
LFA static posture			0.33	0.22	-0.64 (-1.12–-0.17)	-1.39 (-2.22–-0.55)
LFA working hours per day						
>8 hours	398	42				1
≤8 hours	87	9				-0.05 (-0.59–0.49)
≤6 hours	96	10				-0.22 (-0.77–0.33)
≤4 hours	204	22				-0.08 (-0.12–0.04)
≤2 hours	40	4				-0.11 (-0.19–-0.02)
Unknown	119	13				-0.04 (-0.73–0.64)
Self-reported surveys						
SF36						
Physical functioning			41.6	24.8	0.01 (0.00–0.02)	0.01 (-0.01–0.00)
Energy			30.5	17.6	0.02 (0.00–0.03)	0.02 (0.01–0.03)
Health change			37.5	27.8		0.01 (0.00–0.01)
Whitely Index			6.1	3.0	-0.03 (-0.10–0.05)	
Well-being Inventory						
Symptoms			48.4	13.0	0.02 (-0.01– 0.04)	0.01 (-0.01–0.02)
Disability			24.1	4.2	-0.06 (-0.11–-0.01)	-0.05 (-0.10–-0.01)
Workability score			2.5	2.1	-0.47 (-0.61–-0.33)	-0.55 (-0.66–-0.44)

showing that it was difficult to identify relevant claimants with standard regression procedures. The PPV was 62% and the NPV 79%. Eight variables ended up in the standard model: WAS at baseline, work status, WBI disability, wage loss, SF36 energy, SF36 physical functioning, WBI symptoms, and WI.

The weighted model predicted a larger number of improvements compared to the standard model (table 3). The number of predicted cases increased from 16 to 27, ie, from 9% to 14% of the total number of claimants, and was now closer to the percentage of actually observed improvements in the study population (22%). The PPV and NPV were 63% and 82%, respectively. The weighted model contained 11 variables. It included the same variables as the standard model, except for the variable WI. Additionally, the variables LFA static posture, LFA working hours, mental healthcare, and SF36 health change were added. All the VIF scores in the collinearity statistics for the multivariable models were <10, therefore multicollinearity was not assumed. The last two columns of table 1 show the coefficients of the multivariable logit models.

The sensitivity, ie, the model's ability to correctly

Table 2. Descriptive statistics of the study population at baseline of the variables not included in the multivariable models. [HADS= Hospital Anxiety and Depression Scale; SD=standard deviation; SF 36=Short Form Health Survey, 36 items.]

	Study population (N=944)			
	N	%	Mean	SD
Socio-demographics				
Age (years)			51.2	9.0
Gender (female)	476	50		
Educational level				
Low	309	33		
Secondary	399	39		
High	266	28		
Partner (yes)	705	75		
Children (yes)	704	75		
Principal wage earner (yes)	629	67		
Occupational				
Occupational sector				
Finance	127	13		
Government	104	11		
Healthcare	204	22		
Manufacturing	104	11		
Wholesale and retail	120	13		
Other	285	30		
Managerial tasks (yes)	216	23		
Work demands				
Physical	271	29		
Psychological	285	30		
Physical and psychological	388	41		
Health				
Primary diagnosis				
Cardiovascular	96	10		
Mental	233	25		
Musculoskeletal	373	40		
Nervous system	87	9		
Other	155	16		
Comorbidity (yes)	669	71		
Medication use	840	89		
Disability assessment				
Possibility to work (yes)	789	84		
Permanency (yes)	304	32		
List of functional abilities				
Personal functioning			0.08	0.07
Social functioning			0.11	0.12
Adjusting to physical environment			0.11	0.09
Dynamic movement			0.26	0.14
Self-reported surveys				
SF36				
Role limitations due to physical health			6.8	19.9
Role limitations due to emotional problems			33.7	44.2
Emotional well-being			50.3	22.6
Social functioning			53.6	10.4
Pain			37.8	24.9
General health			33.4	17.0
HADS				
Anxiety			9.5	4.8
Depression			9.8	5.0
Well-being Inventory				
Coping			42.5	10.0
Support			56.4	12.3
Stress			37.9	9.5

Table 3. Predictions of the standard and weighted multinominal logit model (test set).

Predicted		Observed		
		Deterioration	No change	Improvement
		N (%)	N (%)	N (%)
Standard model	Deterioration	7 (70)	3 (30)	0 (0)
	No change	25 (16)	100 (62)	36 (22)
	Improvement	0 (0)	6 (38)	10 (62)
Weighted model	Deterioration	8 (57)	6 (43)	0 (0)
	No change	24 (16)	93 (64)	29 (20)
	Improvement	0 (0)	10 (37)	17 (63)

Table 4. Performance measures of the multinominal logit (MNL) models representing sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV).

	MNL model	
	Standard %	Weighted %
Sensitivity	22	37
Specificity	96	93
PPV	62	63
NPV	79	82

detect claimants with an improvement in the WAS, increased from 22% to 37% when we compared the weighted to the standard model (table 4). Both the PPV and NPV of the weighted model were slightly higher as well; the PPV increased from 62% to 63%, and the NPV increased from 79% to 82%. This means that the predictions of the weighted model were correct more often than the predictions of the standard model, although the differences were small.

McNemar's χ^2 was equal to 6.667 and a corresponding P-value of 0.0009. This means that the two models had a different proportion of errors on the test set. The contingency table showed that the number of cases that the weighted model predicted correctly was higher than the number of claimants correctly classified by the standard model. The total number of claimants who were classified differently by the weighted model compared to the standard model was 15, which was sufficiently large to provide accurate P-values for McNemar's test (minimum number is 10) (30).

The results that the weighted model was better at predicting claimants who will experience an improvement in WAS at one-year follow-up for the test set were in line with the results of the training set. In the test set the percentage of claimants identified increased from 9% to 14%.

Discussion

The aims of this study were to (i) predict changes in work ability at one year after approval of the work disability benefit and (ii) explore whether weighted regression procedures could improve the accuracy of predicting claimants with the highest probability of experiencing an improvement in WAS. A minority of 22% of the claimants in our study population experienced an improvement in WAS. Our standard model predicted a relevant improvement in WAS for only 9% of the claimants, while the weighted model predicted this for 14%. However, the PPV of the weighted model did not decrease compared to the standard model. Likewise, the NPV slightly increased. Hence, the weighted model predicted more claimants who will experience a relevant improvement in WAS at one-year follow-up. At the same time, IP can be more certain that the model predicts the correct outcome.

We used a weighted regression model with a linear weight function that assigns larger weights to claimants with a bigger improvement in WAS. Our finding that the weighted model could correctly identify a larger group of individuals with an improvement in WAS in both the training and test sets implies that our weight function could also be of added value in a population that was not used to build the models. However, as the set of possible weight functions is inexhaustible, it could be that there are other weight functions that provide similar or better results than the weight function we have chosen.

The majority of individuals in the study population (63%) did not experience a change in WAS at one-year follow-up. This is in agreement with previous research showing that changes in WAI are small for most individuals, especially for those with longer episodes of sickness absence (18, 31). Determinants of work ability have been reported in several studies. In the present study, work ability at baseline was the strongest predictor in both models. This is in line with previous research showing that, for sick listed workers diagnosed with cancer, WAS at baseline was an important predictor for WAS at one-year follow-up (32). This study also showed an association with wage loss, as we found that individuals with a lower level of wage loss were more likely to experience an improvement in WAS. A higher level of wage loss means more extensive functional limitations, which seems to have a negative effect on work ability at one-year follow-up. This relation was also found for degree of sickness absence and changed WAS at 6- and 12-months follow-up for women on sick leave for ≥60 days (18). Several studies have also found a relation between the WAI and mental and physical conditions, demands at work, individual characteristics and lifestyle (33, 34). These studies did, however, not report measures of diagnostic accuracy (eg, sensitivity and predictive values) of the estimated models.

As pointed out in a recent editorial on prediction models for sickness absence, researchers should be

careful making claims on the accuracy of these models (35). Although the difference between the standard and weighted model in terms of predicting claimants with an improvement in WAS was statistically significant, it was small and it is therefore questionable whether this difference is relevant. However, in the current policies of the SSI, because of the limited capacity to perform IP reassessments and the fact that only a minority of 22% of the individuals actually experienced an improvement in WAS at one-year follow-up, the prediction model may be a relevant tool for identifying the group of claimants with the highest probability of experiencing an improvement in WAS. Our focus was not on predictions at the individual level, but at a population level. Hence, the small differences between the standard and the weighted model are regarded as useful in achieving a more effective allocation of limited occupational health care resources. The weighted model identifying 14% of the claimants, as opposed to 9% with the standard model, with 63% accuracy is considered as a useful auxiliary tool for IP when they plan reassessments. Likewise, in case the model predicts no substantial improvement in WAS at one-year follow-up (which is the case for 86% of the claimants), this could be an indication that for this group of claimants scheduling a reassessment at one-year follow-up has less added value as the NPV is 82%. These probabilities are much higher than the case where the SSI policy is to plan reassessments at random. However, it could be argued that, in other applications, the differences between the two models shown in the present study are too small to be of practical relevance.

We are not aware of any prediction model for future changes in work ability for individuals with a work disability benefit. Previous studies on long-term sickness absence in the general working population have shown that it is difficult to develop prediction models with high prediction accuracy that are relevant in practice. Studies identifying claimants at risk for work disability and long-term sickness absence showed only moderate prediction accuracy (36, 37). Studies on prediction models for individuals with specific chronic diseases such as low-back pain or common mental disorders validated prediction models in terms of PPV and NPV (38–40). Similar to the results of the present study, the NPV of their models were in the range of 74–98% which is considered high. However, they reported PPV of 33–57% which is lower than the PPV of our model (63%).

Strengths and limitations

A strength of the present study is that, by fitting weighted MNL, we are better able to meet practical needs. Non-parametric models offer important advantages because they can focus accuracy on claimants who most likely will experience a change in their entitlement of the work disability benefit. Moreover, by dividing the study sample in a training set to build our prediction models on and a test set to validate the models, we were able to assess the predictive accuracy and generalization of the model.

A further strength is that we combined self-reported questionnaire data with administrative data. This enriches the understanding of a broad range of medical, social, psychological, and work-related factors that can influence future work ability.

Moreover, whereas most studies about predictors of work disability duration and return to work focus on a specific category of diagnoses, our study cohort included a broad range of diseases and disorders. A limitation of our study is that two groups of individuals were excluded from the FORWARD cohort and could therefore also not be included in our study: individuals suffering from severe mental, cognitive, or visual disorders (eg, dementia or psychosis), due to their reduced ability to correctly complete the questionnaires, and individuals diagnosed with cancer.

A study limitation is that the FORWARD cohort questionnaires were not designed to identify the best independent variables for predicting changes in work ability. For instance, own expectation about future changes in work ability were not covered in the questionnaire while the individual's own expectations are important predictors for duration of long-term sick leave and return to work (41, 42). Moreover, the administrative data that we used was not collected for research purposes but rather registered by SSI employees for administration purposes. However, the FORWARD cohort questionnaires are extensive and, by combining them with administrative data, we were able to cover a broad range of potential predictors. A final limitation of this study is our reliance on changes in self-reported work ability. In line with previous studies, we defined an improvement or deterioration in WAS of $\geq$2 points as a relevant change (10, 11, 20). However, it should be investigated if this is also the case for our study population.

Implications for research and practice

Commonly reported outcomes in epidemiological and medical research, such as the incidence of clinical events among a cohort of patients or the response rate in patients taking a certain treatment regimen, are rare events and usually difficult to estimate. Disease predictions can contribute to a wide range of applications, such as risk management, tailored health communication, and decision support systems (43, 44). Weighted analysis could aid these applications by making more accurate predictions of rare events and diseases.

Identification of claimants with a high probability of experiencing an improvement of work ability at one-year follow-up may assist IP during the medical dis-

ability assessment when they need to predict future work ability. This can aid accurate prognosis of work ability and providing suitable interventions to return to work.

To be used in practice, the prediction model needs to be supported by a suitable tool, which is easy to access and interpret for professionals. Future research should focus on the preferable design and content of such a decision support tool. Next, a cost-effectiveness analysis and process evaluation should be performed to determine the added value of the model for IP in making accurate prognoses of work ability.

Concluding remarks

This study showed that, compared to standard MNL models, there are indications that weighted regression procedures can correctly identify more claimants who experience an improvement in WAS. Our findings suggest that a weighted analysis could be an effective method in epidemiology when predicting rare events or diseases. More research is needed to examine the added value of weighted regression procedures in occupational epidemiology.

Acknowledgments

We would like to thank Ms K Bonefaas-Groenewoud for her assistance in processing and preparing the data of the FORWARD study.

Conflicts of interest

IL, HJvR and KHNW are employed at SSI. AJvdB and JRA are shareholders of Amsterdam University Medical Center's spin-off company Evalua Nederland BV. JRA holds a chair in Insurance Medicine on behalf of the SSI.

References

1. Waddell G, Burton AK. Is work good for your health and well-being? London: The Stationary Office; 2006.
2. Viikari-Juntura E, Kausto J, Shiri R, Kaila-Kangas L, Takala EP, Karppinen J et al. Return to work after early part-time sick leave due to musculoskeletal disorders: a randomized controlled trial. Scand J Work Environ Health 2012 Mar;38(2):134–43. https://doi.org/10.5271/sjweh.3258.
3. OECD. Sickness, Disability and Work: Breaking the Barriers: A Synthesis of Findings across OECD Countries. Paris: OECD Publishing; 2010.
4. Krause N, Frank JW, Dasinger LK, Sullivan TJ, Sinclair SJ. Determinants of duration of disability and return-to-work after work-related injury and illness: challenges for future research. Am J Ind Med 2001 Oct;40(4):464–84. https://doi.org/10.1002/ajim.1116.
5. Slebus FG, Sluiter JK, Kuijer PP, Willems JH, Frings-Dresen MH. Work-ability evaluation: a piece of cake or a hard nut to crack? Disabil Rehabil 2007 Aug;29(16):1295–300. https://doi.org/10.1080/09638280600976111.
6. Anner J, Schwegler U, Kunz R, Trezzini B, de Boer W. Evaluation of work disability and the international classification of functioning, disability and health: what to expect and what not. BMC Public Health 2012 Jun;12(1):470. https://doi.org/10.1186/1471-2458-12-470.
7. Louwerse I, Huysmans MA, van Rijssen HJ, van der Beek AJ, Anema JR. Characteristics of individuals receiving disability benefits in the Netherlands and predictors of leaving the disability benefit scheme: a retrospective cohort study with five-year follow-up. BMC Public Health 2018 Jan;18(1):157. https://doi.org/10.1186/s12889-018-5068-7.
8. Kok R, Hoving JL, Verbeek J, Schaafsma FG, van Dijk FJ. Integrating evidence in disability evaluation by social insurance physicians. Scand J Work Environ Health 2011 Nov;37(6):494–501. https://doi.org/10.5271/sjweh.3165.
9. Ilmarinen J. Work ability--a comprehensive concept for occupational health research and prevention. Scand J Work Environ Health 2009 Jan;35(1):1–5. https://doi.org/10.5271/sjweh.1304.
10. Alavinia SM, de Boer AG, van Duivenbooden JC, Frings-Dresen MH, Burdorf A. Determinants of work ability and its predictive value for disability. Occup Med (Lond) 2009 Jan;59(1):32–7. https://doi.org/10.1093/occmed/kqn148.
11. de Boer AG, Verbeek JH, Spelten ER, Uitterhoeve AL, Ansink AC, de Reijke TM et al. Work ability and return-to-work in cancer patients. Br J Cancer 2008 Apr;98(8):1342–7. https://doi.org/10.1038/sj.bjc.6604302.
12. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ 2005 Apr;330(7494):765. https://doi.org/10.1136/bmj.38398.500764.8F.
13. Bright TJ, Wong A, Dhurjati R, Bristow E, Bastian L, Coeytaux RR et al. Effect of clinical decision-support systems: a systematic review. Ann Intern Med 2012 Jul;157(1):29–43. https://doi.org/10.7326/0003-4819-157-1-201207030-00450.
14. King G, Zeng L. Logistic Regression in Rare Events Data. Polit Anal 2001;9(2):137–63. https://doi.org/10.1093/oxfordjournals.pan.a004868.
15. Wei WW. Time Series Analysis. In: Little TD, editor. The Oxford Handbook of Quantitative Methods in Psychology. New York: Oxford University Press; 2013. p. 458–85.
16. Levy PS, Lemeshow S. Sampling of Populations: Methods and Applications. 4th ed. New Jersey: John Wiley & Sons; 2008.
17. Ilmarinen J, Tuomi K. Work ability of aging workers. Scand J Work Environ Health 1992;18(2 Suppl 2):8–10.
18. Ahlstrom L, Grimby-Ekman A, Hagberg M, Dellve L. The

work ability index and single-item question: associations with sick leave, symptoms, and health--a prospective study of women on long-term sick leave. Scand J Work Environ Health 2010 Sep;36(5):404–12. https://doi.org/10.5271/sjweh.2917.

19. El Fassi M, Bocquet V, Majery N, Lair ML, Couffignal S, Mairiaux P. Work ability assessment in a worker population: comparison and determinants of Work Ability Index and Work Ability score. BMC Public Health 2013 Apr;13(1):305. https://doi.org/10.1186/1471-2458-13-305.

20. Boström M, Sluiter JK, Hagberg M. Changes in work situation and work ability in young female and male workers. A prospective cohort study. BMC Public Health 2012 Aug;12(1):694. https://doi.org/10.1186/1471-2458-12-694.

21. Ekman A, Ahlstrand C, Andrén M, Boström M, Dellve L, Eklöf M et al. Ung Vuxen—Basenkät. [Young Adults—Baseline Questionnaire]. Report from the Department of Occupational and Environmental Medicine, Nr 118. Gothenburg: Gothenburg University; 2008.

22. SSI. Het Claimbeoordelings- en Borgsysteem; Een introductie voor belangstellenden. [The Claim Assesment and Monitoring System; an introduction for interested people]. Amsterdam: Dutch Social Security Institute; 2003.

23. Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992 Jun;30(6):473–83. https://doi.org/10.1097/00005650-199206000-00002.

24. Pilowsky I. Dimensions of hypochondriasis. Br J Psychiatry 1967 Jan;113(494):89–93. https://doi.org/10.1192/bjp.113.494.89.

25. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta Psychiatr Scand 1983 Jun;67(6):361–70. https://doi.org/10.1111/j.1600-0447.1983.tb09716.x.

26. Vendrig AA, Schaafsma FG. Reliability and Validity of the Work and Well-Being Inventory (WBI) for Employees. J Occup Rehabil 2018 Jun;28(2):377–90. https://doi.org/10.1007/s10926-017-9729-7.

27. Kaymaz I, McMahon CA. A response surface method based on weighted regression for structural reliability analysis. Probab Eng Mech 2005;20(1):11–7. https://doi.org/10.1016/j.probengmech.2004.05.005.

28. Kennedy P. A Guide to Econometrics. 6th ed. West Sussex: John Wiley & Sons; 2008.

29. Akobeng AK. Understanding diagnostic tests 1: sensitivity, specificity and predictive values. Acta Paediatr 2007 Mar;96(3):338–41. https://doi.org/10.1111/j.1651-2227.2006.00180.x.

30. Siegel SC, Castellan J. Nonparametric statistics for the behavioural sciences. New York: McGraw-Hill; 1988.

31. Feldt T, Hyvönen K, Mäkikangas A, Kinnunen U, Kokko K. Development trajectories of Finnish managers' work ability over a 10-year follow-up period. Scand J Work Environ Health 2009 Jan;35(1):37–47. https://doi.org/10.5271/sjweh.1301.

32. van Muijen P, Duijts SF, Bonefaas-Groenewoud K, van der Beek AJ, Anema JR. Predictors of fatigue and work ability in cancer survivors. Occup Med (Lond) 2017 Dec;67(9):703–11. https://doi.org/10.1093/occmed/kqx165.

33. van den Berg TI, Elders LA, de Zwart BC, Burdorf A. The effects of work-related and individual factors on the Work Ability Index: a systematic review. Occup Environ Med 2009 Apr;66(4):211–20. https://doi.org/10.1136/oem.2008.039883.

34. Radkiewicz P, Widerszal-Bazyl M. Psychometric properties of Work Ability Index in the light of comparative survey study. Int Congr Ser 2005;1280:304–9. https://doi.org/10.1016/j.ics.2005.02.089.

35. Burdorf A. Prevention strategies for sickness absence: sick individuals or sick populations? Scand J Work Environ Health 2019 Mar;45(2):101–2. https://doi.org/10.5271/sjweh.3807

36. Pedersen J, Gerds TA, Bjorner JB, Christensen KB. Prediction of future labour market outcome in a cohort of long-term sick-listed Danes. BMC Public Health 2014 May;14(1):494. https://doi.org/10.1186/1471-2458-14-494.

37. Roelen C, Thorsen S, Heymans M, Twisk J, Bültmann U, Bjørner J. Development and validation of a prediction model for long-term sickness absence based on occupational health survey variables. Disabil Rehabil 2018 Jan;40(2):168–75. https://doi.org/10.1080/09638288.2016.1247471.

38. Dionne CE, Bourbonnais R, Frémont P, Rossignol M, Stock SR, Larocque I. A clinical return-to-work rule for patients with back pain. CMAJ 2005 Jun;172(12):1559–67. https://doi.org/10.1503/cmaj.1041159.

39. Heymans MW, Anema JR, van Buuren S, Knol DL, van Mechelen W, de Vet HC. Return to work in a cohort of low back pain patients: development and validation of a clinical prediction rule. J Occup Rehabil 2009 Jun;19(2):155–65. https://doi.org/10.1007/s10926-009-9166-3.

40. Nieuwenhuijsen K, Verbeek JH, de Boer AG, Blonk RW, van Dijk FJ. Predicting the duration of sickness absence for patients with common mental disorders in occupational health care. Scand J Work Environ Health 2006 Feb;32(1):67–74. https://doi.org/10.5271/sjweh.978.

41. Heijbel B, Josephson M, Jensen I, Stark S, Vingård E. Return to work expectation predicts work in chronic musculoskeletal and behavioral health disorders: prospective study with clinical implications. J Occup Rehabil 2006 Jun;16(2):173–84. https://doi.org/10.1007/s10926-006-9016-5.

42. Sampere M, Gimeno D, Serra C, Plana M, López JC, Martínez JM et al. Return to work expectations of workers on long-term non-work-related sick leave. J Occup Rehabil 2012 Mar;22(1):15–26. https://doi.org/10.1007/s10926-011-9313-5.

43. Cohen EL, Caburnay CA, Luke DA, Rodgers S, Cameron GT, Kreuter MW. Cancer coverage in general-audience and Black newspapers. Health Commun 2008 Sep;23(5):427–35. https://doi.org/10.1080/10410230802342176.

44. Koh HC, Tan G. Data mining applications in healthcare. J Healthc Inf Manag 2005;19(2):64–72.

Received for publication: 27 November 2018