

BMJ Open Identifying risk factors of long sickness absences: a registry-based study using explainable AI methods

Anniina Anttila ,¹ Mikko Nuutinen ,² Riikka-Leena Leskelä ,²
Mark van Gils ,¹ Riitta Sauni ¹

To cite: Anttila A, Nuutinen M, Leskelä R-L, *et al.* Identifying risk factors of long sickness absences: a registry-based study using explainable AI methods. *BMJ Open* 2025;**15**:e101921. doi:10.1136/bmjopen-2025-101921

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (<https://doi.org/10.1136/bmjopen-2025-101921>).

AA and MN contributed equally.

Received 10 March 2025
Accepted 07 October 2025



© Author(s) (or their employer(s)) 2025. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ Group.

¹Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland

²Nordic Healthcare Group Oy, Helsinki, Finland

Correspondence to
Dr Anniina Anttila;
anniina.anttila@finla.fi

ABSTRACT

Objective To identify and explore variable groups and individual predictors of long sickness absences outside of well-known predictors such as service use and previous sickness absence using machine learning, explainable artificial intelligence methods and a submodel approach.

Design Retrospective study of prospectively collected registry data on sickness absences and a questionnaire used in health examinations.

Setting Electronic medical record data of one large occupational health service provider in Finland.

Participants 11 533 employees of various occupations who, between 2011 and 2019, had at least once completed a health questionnaire that could be linked to service usage data and who had not had their initial health check within 1 year before or 3 months after completing the questionnaire.

Primary outcome measures To identify predictors of at least one long sickness absence period (≥ 30 days) during a 2-year follow-up.

Results The highest area under the receiver operating characteristic curve (AUROC) values among the submodel groups were for the sickness absence and service use submodels (0.68–0.74). The AUROC values for the submodels of sociodemographic factors, health habits or diseases data category ranged from 0.55 to 0.67 and from 0.55 to 0.67 for the submodels of questionnaire data. The AUROC value of the ensemble model that combined all submodels was 0.79 (95% CI 0.788 to 0.794).

The most important factors predicting long sickness absences based on the submodels were reported pain, number of symptoms and diseases, body mass index and short sleep duration. Additionally, several work and mental health-related variables increased the risk of long sickness absence.

Conclusions Other variables besides service use and sickness absence increase the accuracy in predicting long sickness absence and providing information for planning interventions that could have a beneficial impact on work disability risk.

INTRODUCTION

Long sickness absences (SA) have been recognised as risk indicators of permanent work disability, mortality and morbidity.¹ Work disability also causes disturbances in the workflow and productivity of the workplace,

STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ The study was conducted on a large, real-world dataset comprising employees from various industries.
- ⇒ The dataset contained a rich and diverse set of categories such as service use and questions on health-related topics, such as eating habits, exercise habits, diseases and symptoms and on work-related themes, such as physical and psychosocial risk factors.
- ⇒ Sickness absence (SA) data were available from the first day of absence.
- ⇒ The machine learning submodels were trained for different categories to estimate the prediction accuracy of the submodels and discover the most predictive variables outside the well-known predictors of long SA.
- ⇒ A limitation was that the study was performed on the data from only one occupational health service provider in Finland.

besides the financial, social and health problems for the individual, and it has a significant economic impact on the social security system. The prevention of work disability has been recognised as an important objective across OECD countries in view of the ageing of the working population and the aim of prolonging working careers.²

Interventions in occupational health services that target patients on sick leave can promote return to work, as numerous studies have shown.³ Earlier interventions by occupational health services before the onset of work disability can also reduce SA rates and health-related retirement.⁴ It is crucial that occupational health professionals are able to recognise patients at risk of recurrent or long SA and disability pensions as early as possible. Occupational health intervention programmes can be both cost-saving and more effective than the usual occupational healthcare when targeted at selected high-risk employees.⁵

The capabilities of data analysis methods for risk assessment have evolved with the constant improvement of artificial intelligence (AI) and machine learning (ML) technologies and especially with the increasing availability of large and rich datasets combined with powerful computing resources. ML models have been extensively studied to predict various health outcomes.^{6–8} However, research on the implementation of these capabilities in an occupational health setting with work disability as the outcome is still limited.

When an ML model is trained using a learning algorithm, its parameters iterate towards an efficient mathematical association between input variables and given outcomes, based on available example data. Based on the successfully established associative model, risk factors for SA and suitable targets for possible interventions can be discovered by using explainable AI (XAI) methods. Age, sex, previous SA, service use and self-rated health have been shown to be strong predictors for identifying individuals at increased risk of work disability.^{9–15} However, they are not risk factors that can be acted on with targeted occupational health interventions. They should instead be considered variables for various underlying causes that may not independently be strong predictors of long-term SA but may be important and understandable factors affecting an individual's work ability level. Therefore, the ML model will most likely assign the largest weights to the strong predictors if all variables, both the underlying causes and the strong predictors, are entered into the same ML model. The model may give a good prediction, but the prediction has limited relevant practical value in the sense that factors such as previous service use or SA cannot be acted on. Identifying the associations between SAs and modifiable factors such as health behaviour can help professionals select appropriate interventions to support work ability.

We used ML submodels in this study for different data categories and XAI methods, such as Shapley values (SHAP), partial dependence plots (PDPs) and surrogate models,^{16–18} to assess the predictive capability of variable groups from different data categories and individual variables outside the well-known predictors of long SA. Our research questions were, “What is the accuracy of ML models using different variable groups from various data categories of occupational healthcare data for predicting long-term SA” and “What are the most important variables from different data categories, besides previous SA, service use and general self-rated health?”. The novelty of the study lies in understanding the associations between long SA and individual predictors. The study also provides new information about prediction accuracy that can be obtained with different categories of data.

METHODS

Study population

The present study is a retrospective registry study. We had access to the database of one occupational health service

provider in Finland (Finla). The study participants were the employees of several companies who used Finla's occupational healthcare services during the years 2009–2021 and completed a health questionnaire administered by Finla at least once during the years 2011–2019. A total of 18 840 questionnaires could be linked to service usage data (eg, occupational healthcare visits). Questionnaires whose response time was not recorded were excluded (N=334). Patients whose initial health check was within 1 year before or 3 months after the questionnaire was completed were excluded (N=647) to ensure availability of data on previous service use. The COVID-19 pandemic would have affected SA during the follow-up time, so questionnaire responses later than 18 March 2018 (N=3992) were also excluded. The total number of completed questionnaires with service use data after these adjustments was N=14 514. Online supplemental figure S1 presents a flow chart of how participants were selected for the analyses. The study protocol is available in the online supplemental material.

Data

The main source of predictor variables was the health questionnaire administered by Finla used in initial and periodic health monitoring. It contained questions on both health-related topics, such as eating habits, exercise habits, diseases and symptoms, and on work-related themes, such as physical and psychosocial risk factors. These data were supplemented by variables obtained from the electronic medical records of Finla containing occupational health service usage, diagnoses and SA episodes.

Outcome

The predicted outcome of the models was the occurrence of one or more long (>30 days) SA episodes during a 2-year follow-up time that began on the questionnaire response date. All service use variables and previous SA days were calculated from the period of 1 year before questionnaire completion (online supplemental figure S2).

Additionally, repetitive, short-term SAs were selected as an alternative outcome in the sensitivity analysis to assess the models' robustness. Repetitive, short-term SAs were defined as more than five short (1–10 days) SA episodes during a 2-year follow-up time.

Submodels

The aim of the submodel-specific analyses was to measure the performance of methods that use variable groups from different data categories to predict long-term SA and to explore the individual predictors in more detail. We based our study on training and analysing a series of 12 ML submodels from three data categories. (1) Five submodels were trained from the sociodemographic factors, health habits and diseases data category: (1.1) demography, (1.2) job description, (1.3) measurements, (1.4) health habits and (1.5) diseases and symptoms submodels. (2) Four submodels were trained from the

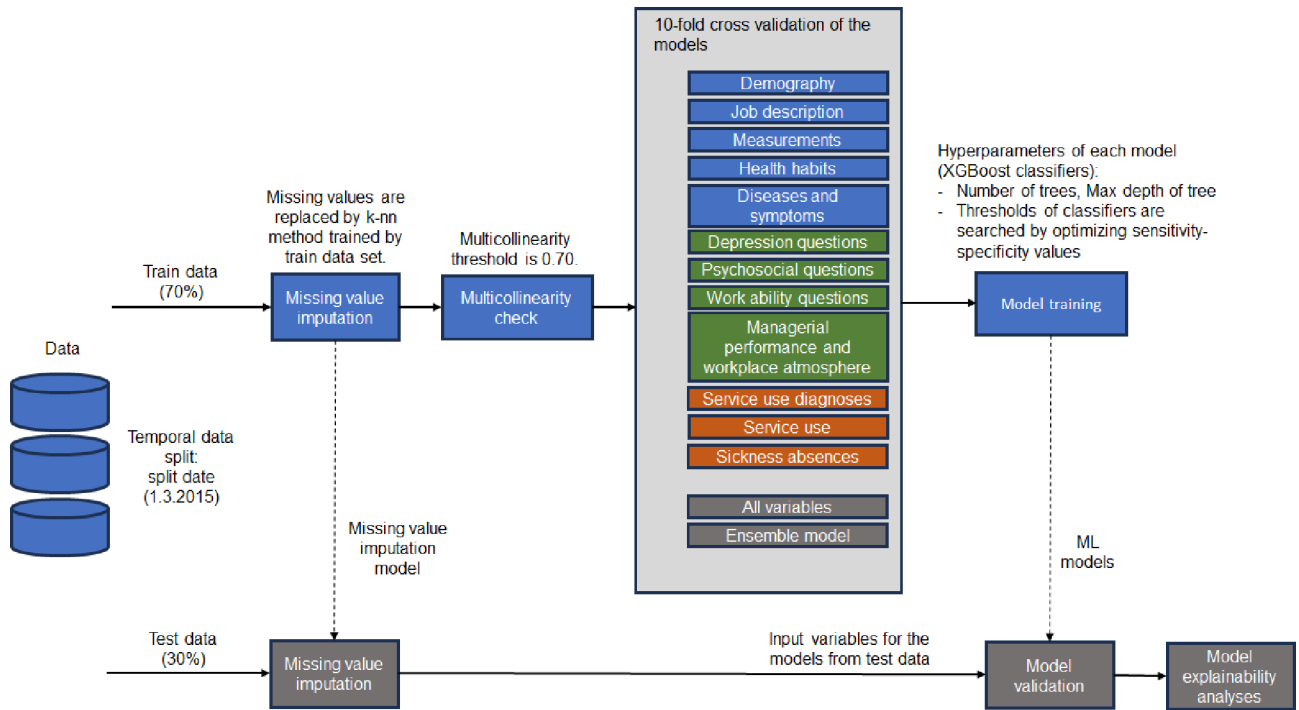


Figure 1 Data flow of machine learning (ML) model training and evaluating pipeline in an occupational health cohort in 2011–2019.

working conditions and mental health data category: (2.1) depression questions, (2.2) psychosocial questions, (2.3) work ability and (2.4) managerial performance and workplace atmosphere submodels. (3) Three submodels were trained from the service usage data category: (3.1) service use diagnoses, (3.2) service use and (3.3) SA submodels. The research questions of the study focus on the submodels from data categories (1) and (2). Data category (3) was included as a reference for model performance. online supplemental table S1 lists the selected variables of each submodel and presents descriptive statistics of the variables. The online supplemental section ‘Collected variables’ presents the definitions and data sources of all variables of each submodel.

Statistical methods

The data set was first divided into two distinct data folds in the ratio of 70% for training (before index date 1.3.2015) and 30% for testing (after index date 1 March 2015) (figure 1 and online supplemental figure S3). A single patient could have completed several questionnaires; thus, possible data leakage was mitigated by removing all patients from the test data ($n=1824$) who already had a questionnaire in the training data. The training data contained 8874 separate questionnaires (7912 patients). The test data contained 3816 separate questionnaires (3621 patients). The training fold was used to select hyperparameters and to train submodels and the test fold for validation of the submodels. Preprocessing steps before the training of the submodels were missing value imputation and multicollinearity check. Missing values were imputed by using k-nearest

neighbour approach.¹⁹ A multicollinearity check was processed separately for each submodel. One of the variables was removed from the submodel if a correlation between input variable pairs was greater than 0.7. The removed variable’s selection was based on the richness of the information the variables held. For example, if we had a variable pair of ‘smoking (yes/no)’ and ‘smoking years’, we selected the variable ‘smoking years’ for the modelling. That is, we tried to maximise the information of the selected variables. Online supplemental table S1 shows the variables excluded from the model training due to multicollinearity.

All hyperparameters were searched from the training data by using the k-fold cross validation method ($k=10$).²⁰ The gradient boosting (XGBoost) algorithm¹⁸ was used as a base learner for the submodels. Two general models f_{all} and f_{ens} were developed for predicting long SA in addition to the 12 submodels. The general models were trained to act as a reference, representing the performance when all available variables are included in the modelling process. The model f_{all} was trained by using variables from all data categories. The input variables of model f_{ens} (ensemble model) were the weighted outcome values from the submodels. The weighting factors were the relative accuracy values of the submodels for training data—in other words, variables that showed better predictive capability of long SA in the training fold were given stronger weight in proportion to their prediction accuracy. Model explainability analysis^{16–18} was performed separately for each submodel. First, the input variables of each submodel were ranked based on the absolute Shapley values. The



surrogate models and PDPs were then used to discover insights from the important variables.

The means and frequencies of the baseline characteristics were used to examine the differences between the patients with and without long SA periods. The χ^2 and Mann-Whitney U test were used to test the differences between the groups. A $p < 0.05$ was considered statistically significant. The area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, positive predictive value (PPV/precision), negative predictive value (NPV), F1 score and the area under the precision recall curve (AUPRC) were used as performance metrics for the submodels and general models. The threshold values for the classification metrics of patients belonging to the positive group were searched by balancing the sensitivity and specificity values of the training data. For performance metrics of testing data, 95% CIs were derived by non-parametric bootstrap (1000 samples). For performance metrics of training data, 95% CIs were derived from the samples of k-fold cross-validation. The online supplemental section 'Data preprocessing' presents a detailed description of the data flow, data preprocessing, model training and evaluating steps.

Patient and public involvement

The patients and/or the public were not involved in the design, conduct, reporting or dissemination plans of this research.

RESULTS

The study comprised 11 533 patients and 14 514 separate questionnaires. Table 1 presents the properties of the patients with or without long SA periods during follow-up. The values were calculated according to the first questionnaire if an individual patient had answered more than one questionnaire. The average age was 43.5 years (SD 11.4; range 16.8–69.8). Of the respondents, 42.5% (N=4902) were female. Altogether, 17.4% worked in a supervisory and 43.3% in a blue-collar position. Furthermore, 23.5% of them did shift work and 6.0% did night work. The most common self-reported diseases were insomnia (15.8%), musculoskeletal disease (14.0%) and hypertension (13.2%). There were statistically significant differences between the groups with and without long SA in all background variables except exercise habits. Online

Table 1 Descriptive statistics of background variables in a Finnish occupational health cohort in 2011–2019

Item name	All	NO long SA periods	Long SA periods	P value
Patients, n	11 533	10 819	714	
Age (years), mean (SD)	43.5 (11.4)	43.3 (11.4)	47.0 (10.5)	<0.001
Sex (female), n (%)	4902 (42.5)	4485 (41.5)	417 (58.4)	<0.001
BMI (kg/m ²), mean (SD)	26.6 (4.5)	26.5 (4.5)	27.7 (5.1)	<0.001
Smoker, n (%)	3103 (26.9)	2885 (26.7)	218 (30.5)	0.026
Exercise habits points (0–12), mean (SD)	5.1 (2.2)	5.1 (2.2)	5.1 (2.3)	0.809
Supervisor, n (%)	2004 (17.4)	1924 (17.8)	80 (11.2)	<0.001
White-collar worker, n (%)	4995 (43.3)	4811 (44.5)	184 (25.8)	<0.001
Blue-collar worker, n (%)	6152 (53.3)	5660 (52.3)	492 (68.9)	<0.001
Shift work, n (%)	2711 (23.5)	2462 (22.8)	249 (34.9)	<0.001
Night work, n (%)	691 (6.0)	635 (5.9)	56 (7.8)	0.034
Does your disease hinder coping at work, n (%)	1119 (9.7)	939 (8.7)	180 (25.2)	<0.001
Asthma, pulmonary disease, n (%)	771 (6.7)	702 (6.5)	69 (9.7)	0.002
Diabetes, n (%)	427 (3.7)	376 (3.5)	51 (7.1)	<0.001
Cardiovascular disease, n (%)	635 (5.5)	563 (5.2)	72 (10.1)	<0.001
Hypertension, n (%)	1527 (13.2)	1395 (12.9)	132 (18.5)	<0.001
Musculoskeletal disease, n (%)	1616 (14.0)	1404 (13.0)	212 (29.7)	<0.001
Common mental disease, n (%)	723 (6.3)	628 (5.8)	95 (13.3)	<0.001
Insomnia, tiredness, n (%)	1817 (15.8)	1648 (15.2)	169 (23.7)	<0.001
Cancer, n (%)	249 (2.2)	212 (2.0)	37 (5.2)	<0.001
SA dg mental and behavioural, n (%)	307 (2.7)	257 (2.4)	50 (7.0)	<0.001
SA dg musculoskeletal, n (%)	1660 (14.4)	1387 (12.8)	273 (38.2)	<0.001
SA days sum, mean (SD)	6.7 (19.6)	5.5 (15.8)	25.5 (44.8)	<0.001

P values are calculated using the χ^2 or Mann-Whitney U test.
BMI, body mass index; SA, sickness absence.

Table 2 Performance of machine learning models for predicting long sickness absence period (test data) in a Finnish occupational health cohort in 2011–2019

Sociodemographic factors, health habits and diseases					
Model	Demography	Job description	Measurements	Health habits	Diseases and symptoms
AUROC	0.663	0.616	0.553	0.562	0.671
Sensitivity	0.659	0.735	0.456	0.597	0.668
Specificity	0.622	0.426	0.622	0.480	0.617
PPV	0.099	0.075	0.070	0.067	0.099
NPV	0.967	0.962	0.948	0.950	0.967
F1 score	0.172	0.135	0.122	0.121	0.172
AUPRC	0.137	0.091	0.071	0.083	0.128
Working conditions and mental health					
Model	Depression questions	Psychosocial questions	Work ability questions	Managerial performance and workplace atmosphere	
AUROC	0.598	0.626	0.668	0.554	
Sensitivity	0.491	0.646	0.566	0.473	
Specificity	0.650	0.564	0.693	0.628	
PPV	0.081	0.085	0.104	0.074	
NPV	0.953	0.962	0.962	0.950	
F1 score	0.139	0.151	0.176	0.128	
AUPRC	0.097	0.101	0.134	0.073	
Service usage (reference)				Full model	
Model	Service use diagnoses	Service use	Sickness absences	All variables	Ensemble
AUROC	0.709	0.738	0.677	0.766	0.790
Sensitivity	0.681	0.597	0.429	0.549	0.650
Specificity	0.650	0.732	0.840	0.812	0.759
PPV	0.109	0.123	0.145	0.155	0.145
NPV	0.970	0.967	0.959	0.966	0.972
F1 score	0.188	0.204	0.217	0.242	0.237
AUPRC	0.169	0.178	0.160	0.202	0.230

The threshold values for the classification metrics* of patients belonging to the positive group of long sickness absence period were automatically selected by balancing the sensitivity and specificity of the training data.

*Sensitivity, specificity, PPV, NPV, F1-score.

AUPRC, area under the precision recall curve; AUROC, area under the receiver operating characteristic curve; Ensemble, prediction model, where the input values were weighted according to the relative accuracy values of the submodels in the training data; NPV, negative predictive value; PPV, positive predictive value.

supplemental table S2 presents subject demographics properties for training and test data sets.

ML models performance

Table 2 presents AUROC, sensitivity, specificity, PPV, NPV, F1 score and AUPRC values of the submodels and general models for the test data set. The AUROC values of the submodels within the Sociodemographic factors, health habits and diseases data category ranged from 0.553 (measurements submodel) to 0.671 (diseases and

symptoms submodel) and for the submodels of the questionnaire data category from 0.554 (managerial performance and workplace atmosphere submodel) to 0.668 (work ability submodel). The AUROC values of the reference submodels of the service use data category were from 0.677 (SA submodel) to 0.74 (service use submodel). For the full models, the AUROC value of the ensemble model was 0.790 and of the model with all variables 0.768. Figure 2 presents the receiver operating characteristic

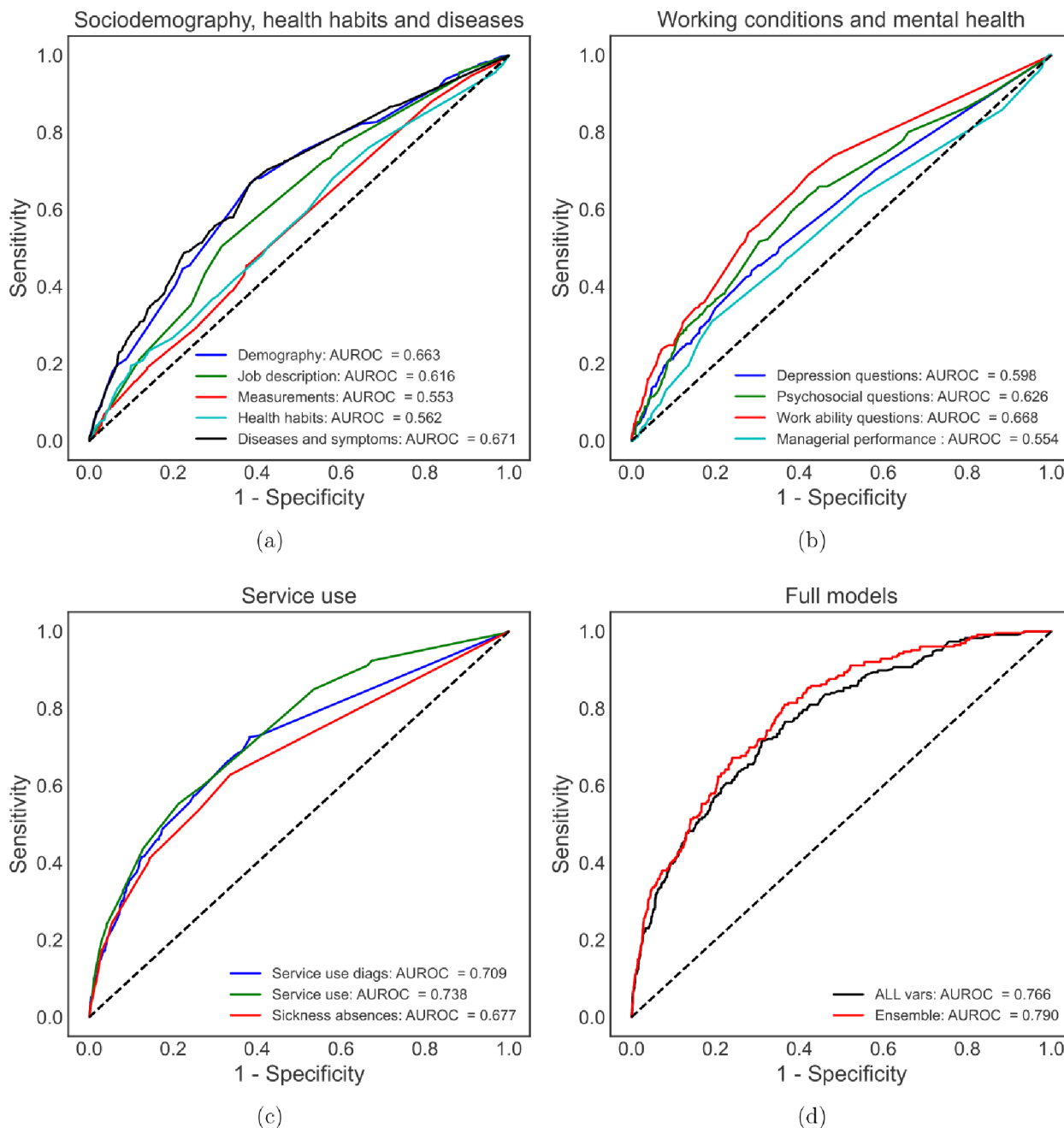


Figure 2 Receiver operating characteristics curves for the prediction of long sickness absence periods in a Finnish occupational health cohort in 2011–2019. The diagonal line indicates no discrimination above chance. (a) Submodels from the data category sociodemographic factors, health habits and diseases; (b) Submodels from the data category working conditions and mental health; (c) Submodels from the data category service usage (reference model) and (d) full models. AUROC, area under receiver operating curve; Ensemble, prediction model, where the input values were weighted according to the relative accuracy values of the submodels in the training data.

curves for the models. Online supplemental table S3 presents the training data's performance values. Online supplemental figures S4–S7 present the correlation matrix of the variables of each data category. Online supplemental figure S8 presents the distributions of the probabilities of long SA predicted by the models for the groups with or without long SA periods at the follow-up. Online supplemental figure S9 presents the distributions of the cumulative SA days over the 1 year of follow-up between patients who were predicted to be at a low or high risk of

a long SA period. Online supplemental figures S10–S18 present the models' XAI analyses. **Figure 3** presents the sum of absolute Shapley values for the variables of each submodel.

The results of the sensitivity, specificity, PPV and NPV depend on the threshold value chosen for the classifiers. Tuning this threshold is an optimisation problem that depends on whether one wishes to minimise false positives, maximise the detection of true positives or strike a balance in between. The thresholds in this study were

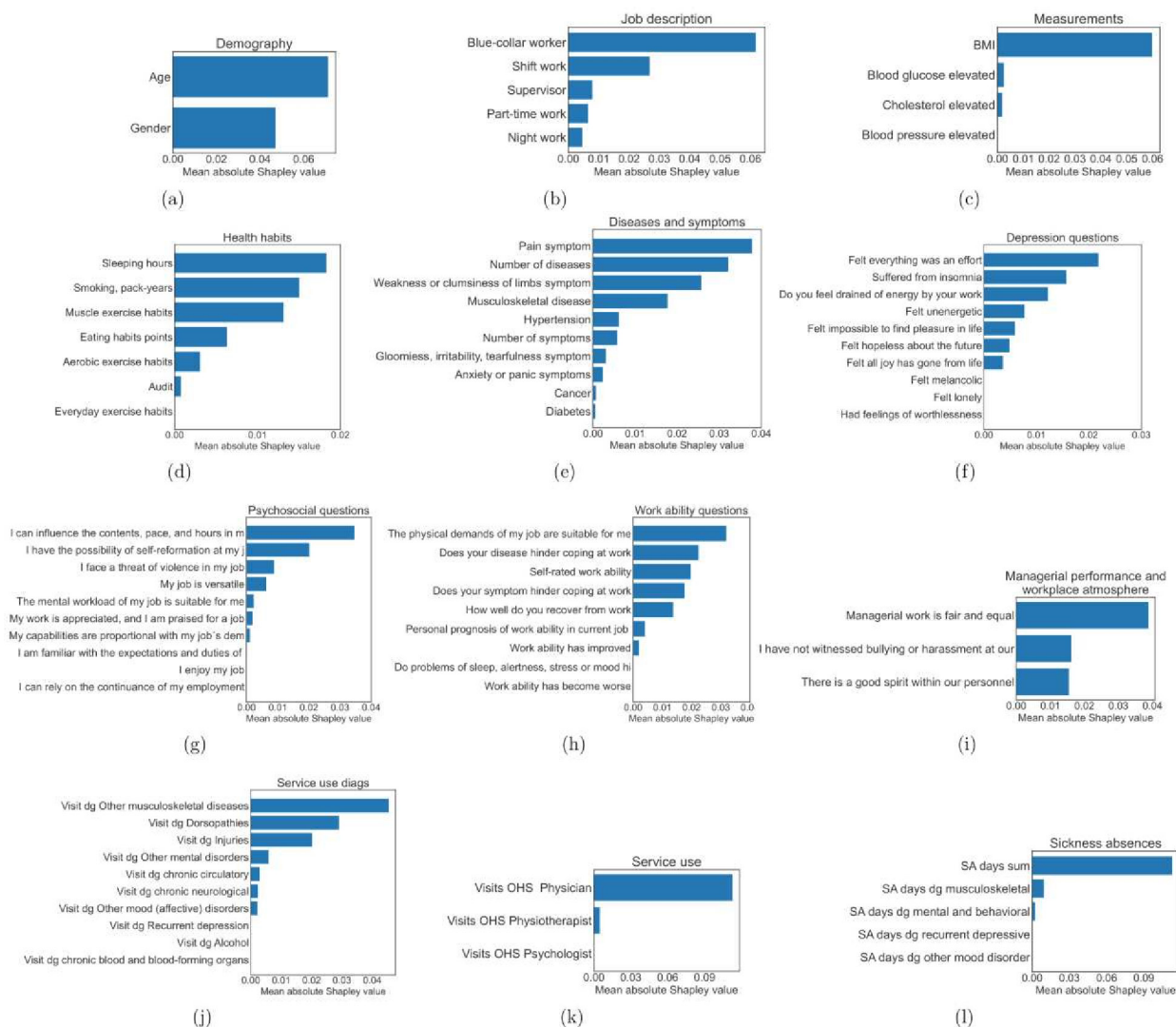


Figure 3 Sum of absolute Shapley values for the variables of the submodels in a Finnish occupational health cohort in 2011–2019. BMI, body mass index; SA, sickness absence; OHS, Occupational Health Services.

chosen automatically by balancing sensitivity and specificity in the training data. Online supplemental table S3 presents the threshold values. The sensitivity of the ensemble model with the selected threshold value (0.383) for the test data was 0.655, indicating that the model found 66% of patients who had a long SA period. The specificity of the ensemble model was 0.733, indicating that 73% of patients classified as low risk did not have a long SA period.

The ensemble model’s NPV was 0.972, and its PPV was 0.145. The NPV represents the share of predictions assigned to the negative group whose risk did not materialise, while the PPV represents the share of positive predictions in which the risk did materialise. Note that PPV and NPV are heavily influenced by the different classes’ prevalence in the datasets. NPV will easily be high (and PPV difficult to get high) and vice versa if there are many negative class values in the dataset (as in our case).

One point for consideration is the profile of the patients that the model classified as high risk. They may indeed have an elevated risk that simply did not materialise

according to the study’s outcome criterion, which was a rather strict definition requiring a long, uninterrupted SA spell. Online supplemental figure S9 presents an analysis comparing the cumulative SA days over the 1-year follow-up between patients predicted to be at high risk and those whose predicted risk was below the threshold. This analysis showed that, especially for the ensemble model, the distribution of SA days differed markedly between the predicted high-risk group and the non-risk group.

Online supplemental table S4 presents the results of the modelling performance’s sensitivity analyses with repetitive, short SA episodes as the outcome. The models’ performance was similar to that observed in the case of long SA episodes.

DISCUSSION

This study developed several ML submodels and two general models for predicting long SAs. An individual submodel comprised a group of variables from a specific



category, such as sociodemographic characteristics, health habits, diseases and questionnaires on working conditions and mental health (depressive symptoms, psychosocial workload, managerial performance and workplace atmosphere, and work ability). The submodels' performance differed. Information on submodel performance can be leveraged to determine the relative importance of the various data sources in occupational health services when developing predictive models. Furthermore, we managed to identify several individual predictors with possible relevant thresholds from the submodels (online supplemental figures S10–18). In addition to strong predictors that have been constantly identified in previous studies, such as age, sex, service use and previous SA days, we found several predictors from the groups 'job description' and 'diseases' and from answers to questionnaires on working conditions and mental health.

The AUROC value of the general ensemble model was 0.79. Among the submodels, the highest AUROC value (0.74) was for the 'service use' submodel. However, the predictive accuracy increases when the variables from the other submodels are included with the service use model in the ensemble model, indicating that the other models have predictive power that is complementary to that of the service use. The performance of the general ensemble model can be considered good compared with the performance values presented in previous studies, where the AUROC value for general models for predicting SA has been between 0.67 and 0.79.^{13 21–26} Of course, the performance values of different studies should be compared with caution, because the data used and what is predicted affect the performance that can be achieved. Additionally, according to our knowledge, previous studies have not always clearly divided the data into training and validation parts, as recommended in the development of ML models.^{13 27} Several previous studies have followed traditional statistical model analysis methods in which the focus is not necessarily on proving the generalisability of the models but rather on proving the fit of the model and data.^{28–31}

The analysis of the submodels indicated that among the work-related factors, the risk of long SA was particularly associated with non-regular work (part-time, shift work). The possibility of altering an employee's working time arrangements, such as changing from shift work to a fixed day shift or providing full-time employment instead of part-time work, varies between workplaces and is not always possible. However, employees in non-regular work should be recognised as a risk group who may need support for maintaining work ability. Other studies have also found that organising shift work in physiologically optimal scheduling (≤ 3 consecutive night shifts, shift intervals of ≥ 11 hours, ≤ 9 hours shift duration) would benefit the welfare of the whole working community by reducing the risk of injuries, SA and possibly breast cancer.^{32–34} Therefore, changes to working time solutions should be encouraged by occupational health professionals if a suboptimal shift work pattern is identified as

the cause for an increased risk of work disability either at an individual or the workplace level.

Higher risk was particularly associated with the working conditions and mental health data category with the subjective feeling of being unable to influence the content of one's work or of not having the possibility of self-reformation, the physical load of the work not seeming appropriate, some illness hindering the work or that the supervisor's actions seemed unfair. Of these, the discrepancy between an employee's health status and the physical or mental workload can be partly reduced by health improvements through eg, medical treatment or rehabilitation. However, most of these factors can only be improved by measures taken at the workplace. Interventions aimed at reducing work disability risk at an individual level are likely to be most effective when they include both elements for improving health status and, when possible, adjusting the work demands.³⁵

The most important variables in the measurements submodel and the health habits submodel were body mass index (BMI), sleep, smoking, muscle training and eating habits. The association between obesity and the risk of work disability is well-known.³⁶ The risk of long SA in our study increased already at a BMI threshold of 27 kg/m^2 . The surrogate model (online supplemental figure S12) also found the risk groups based on BMI value ranges of 22–27, 27–32, >32 , which are close to the BMI risk groups defined by the WHO (18.5–25, 25–30, 30–35, >35).³⁷

Interventions focusing on promoting healthy lifestyles are often initiated at primary healthcare visits and can have a major impact on health parameters.³⁸ It has also been shown that extending such interventions to the workplace level can have a positive impact on somatic and mental diseases as well as on work disability and can be cost effective when considering the expenses of work disability.^{39–41} Improving health habits such as diet and exercise can also reduce the risk of musculoskeletal disorders and hypertension, which were important factors in the diseases and symptoms submodel for predicting a higher risk of long-term SA.

We also performed sensitivity analyses with repetitive, short SAs as an alternative outcome (online supplemental table S4), in addition to our primary outcome of long-term SA, to study the models' robustness. The models' predictive power for repetitive, short SAs was similar to long SAs, which implies that the data categories in our models can be used to predict different patterns of SA.

This study has several strengths. Our real-world data included screening questionnaires that covered a broad spectrum of both work-related and health-related themes and could be linked with data from occupational health medical records. The data represented various industries' workplaces and included SA data from the first day of absence. To our knowledge, the ML and (XAI) methods we applied have not been used similarly in an occupational health setting with work disability as an outcome. Our submodel-based approach was sensitive to detecting not only the strong predictive variables, such as service

utilisation and prior SA, but also other categories of variables, some of which may also be modifiable, unlike the strong predictors. Furthermore, the submodel-based approach measured the performance values of variable groups from different data categories to predict long-term SA. This is important information when analysing the feasibility of model development and important data sources' availability.

The present analyses have some limitations. One limitation is that the ML models were trained on data of one occupational health service provider in Finland. External validation with different patient groups is required to gain a better understanding and test the transferability and generalisability of the models. Generalisation in Finland can be expected to some extent, because the distribution of the different industries of our patient group was close to the proportions of industries in Finland nationwide (online supplemental figure S19). The second limitation is related to the functionality of conventional supervised learning. ML models are effective in finding patterns (associative relationships) in large data sets. We analysed the data in this study to identify patterns that characterise occupational health service clients at risk of long SA periods. While a variable may be significant in the prediction of an outcome, this does not necessarily imply a causal relationship between the variable and the outcome.

CONCLUSIONS

Previous service use and SAs are strong predictors of work disability and form the basis of a general prediction model for long SA. However, identifying these predictors does not in itself provide occupational health professionals with concrete targets for interventions. This study used submodels and XAI methods to identify individual associations between other variables and SAs. These variables are more concrete and understandable risk factors than previous service use and SA and can, therefore, be used to plan interventions that could have an impact on the risk of work disability.

The study's results demonstrated that ML methods can be successfully applied in using occupational health data to identify employees with an increased SA risk. Reducing sick leave will increase workers' well-being and the productivity of companies and society as a whole.

Acknowledgements We want to thank Anu Pekki for her perceptive comments at different stages of this work and Yrjänä Hynninen for his input at the initial phases of this study.

Contributors AA, MN, R-LL, MvG and RS participated in planning the study and interpreting the results. MN conducted the statistical analyses. AA and MN wrote the first draft of the manuscript, and all authors commented on and approved the final manuscript as submitted. AA and MN contributed equally to this paper. AA is the guarantor. AI methods were essential in this study, as they were used in data analyses as described in the manuscript under 'Statistical Methods' and in more detail in the supplement material. AI was not used in producing text or image content.

Funding This study was funded by the Finnish Work Environment Fund (project no: 220127).

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval This study was approved by Finnish data authority (Findata) (THL/1850/14.02.00/2022). The Finnish National Board on Research Integrity TENK guidelines state that ethics approval is not required for retrospective registry studies in human sciences, unless the research threatens the safety of the participants or researchers (<https://tenk.fi/en/advice-and-materials/guidelines-ethical-review-humansciences>). This study used solely secondary data retrieved from registries. Patient information was pseudonymised, and only the members of the research team had access to process and analyse data in a secure, closed environment. Ethics approval was therefore not required for this study. Further inquiries may be directed to Dean Seppo Parkkila, Faculty of Medicine and Health Technology seppo.parkkila@tuni.fi.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <https://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Anniina Anttila <https://orcid.org/0009-0006-5234-0332>
 Mikko Nuutinen <https://orcid.org/0000-0002-7429-3710>
 Riikka-Leena Leskelä <https://orcid.org/0000-0002-9255-2958>
 Mark van Gils <https://orcid.org/0000-0002-0029-1771>
 Riitta Sauni <https://orcid.org/0000-0002-9808-3638>

REFERENCES

- Wallman T, Wedel H, Palmer E, *et al*. Sick-leave track record and other potential predictors of a disability pension. A population based study of 8,218 men and women followed for 16 years. *BMC Public Health* 2009;9.
- Sickness O. Disability and work: breaking the barriers: a synthesis of findings across OECD countries. *OECD* 2010.
- van Vilsteren M, van Oostrom SH, de Vet HCW, *et al*. Workplace interventions to prevent work disability in workers on sick leave. *Cochrane Database Syst Rev* 2015;2015.
- de Boer AGEM, van Beek J-C, Durinck J, *et al*. An occupational health intervention programme for workers at risk for early retirement; a randomised controlled trial. *Occup Environ Med* 2004;61:924-9.
- Taimela S, Justén S, Aronen P, *et al*. An occupational health intervention programme for workers at high risk for sickness absence. Cost effectiveness analysis based on a randomised controlled trial. *Occup Environ Med* 2008;65:242-8.
- Kourou K, Manikis G, Mylonas E, *et al*. Personalized prediction of one-year mental health deterioration using adaptive learning algorithms: a multicenter breast cancer prospective study. *Sci Rep* 2023;13:7059.
- Nuutinen M, Hiltunen A-M, Korhonen S, *et al*. Aid of a machine learning algorithm can improve clinician predictions of patient quality of life during breast cancer treatments. *Health Technol* 2023;13:229-44.
- Nuutinen M, Haukka J, Virkkula P, *et al*. Using machine learning for the personalised prediction of revision endoscopic sinus surgery. *PLoS One* 2022;17.
- Airaksinen J, Jokela M, Virtanen M, *et al*. Prediction of long-term absence due to sickness in employees: development and validation of a multifactorial risk score in two cohort studies. *Scand J Work Environ Health* 2018;44:274-82.



- 10 Roelen CAM, Bultmann U, Stapelfeldt CM, *et al.* Multicentre validation of frequent sickness absence predictions: table 1. *OCCMED* 2016;66:69–71.
- 11 Roelen CAM, Heymans MW, Twisk JWR, *et al.* Health measures in prediction models for high sickness absence: single-item self-rated health versus multi-item SF-12. *Eur J Public Health* 2015;25:668–72.
- 12 Roelen CA, van Rhenen W, Groothoff JW, *et al.* The development and validation of two prediction models to identify employees at risk of high sickness absence. *Eur J Public Health* 2013;23:128–33.
- 13 Nyberg ST, Elovainio M, Pentti J, *et al.* Predicting long-term sickness absence with employee questionnaires and administrative records: a prospective cohort study of hospital employees. *Scand J Work Environ Health* 2023;49:610–20.
- 14 LoMartire R, Dahlström Ö, Björk M, *et al.* Predictors of sickness absence in a clinical population with chronic pain. *J Pain* 2021;22:1180–94.
- 15 Mortensen J, Dich N, Lange T, *et al.* Job strain and informal caregiving as predictors of long-term sickness absence: a longitudinal multi-cohort study. *Scand J Work Environ Health* 2017;43:5–14.
- 16 Craven M, Shavlik J. Extracting tree-structured representations of trained networks. *Advances in Neural Information Processing Systems*, MIT Press; 1995 Available: <https://proceedings.neurips.cc/paper/1995/hash/45f31d16b1058d586fc3be7207b58053-Abstract.html>
- 17 Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, Curran Associates, Inc; 2017 Available: <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
- 18 Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Statist* 2001;29:1189–232.
- 19 Troyanskaya O, Cantor M, Sherlock G, *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* 2001;17:520–5.
- 20 Hastie T, Friedman J, Tibshirani R. Model assessment and selection. In: *The Elements of Statistical Learning, in Springer Series in Statistics*. New York, NY: Springer, 2001: 193–224.
- 21 Holm J, Frumento P, Almondo G, *et al.* Predicting the duration of sickness absence due to knee osteoarthritis: a prognostic model developed in a population-based cohort in Sweden. *BMC Musculoskelet Disord* 2021;22:603.
- 22 van Hoffen MFA, Norder G, Twisk JWR, *et al.* External validation of a prediction model and decision tree for sickness absence due to mental disorders. *Int Arch Occup Environ Health* 2020;93:1007–12.
- 23 van Hoffen MFA, Norder G, Twisk JWR, *et al.* Development of prediction models for sickness absence due to mental disorders in the general working population. *J Occup Rehabil* 2020;30:308–17.
- 24 Louwse I, van Rijssen HJ, Huysmans MA, *et al.* Predicting long-term sickness absence and identifying subgroups among individuals without an employment contract. *J Occup Rehabil* 2020;30:371–80.
- 25 van der Burg LRA, van Kuijk SMJ, ter Wee MM, *et al.* Long-term sickness absence in a working population: development and validation of a risk prediction model in a large Dutch prospective cohort. *BMC Public Health* 2020;20.
- 26 Gémes K, Holm J, Frumento P, *et al.* A prognostic model for predicting the duration of 20,049 sickness absence spells due to shoulder lesions in a population-based cohort in Sweden. *PLoS One* 2023;18.
- 27 Tohka J, van Gils M. Evaluation of machine learning algorithms for health and wellness applications: a tutorial. *Comput Biol Med* 2021;132:104324.
- 28 Real E, Jover L, Verdaguer R, *et al.* Factors associated with long-term sickness absence due to mental disorders: a cohort study of 7,112 patients during the spanish economic crisis. *PLoS One* 2016;11.
- 29 Roelen CAM, van Hoffen MFA, Waage S, *et al.* Psychosocial work environment and mental health-related long-term sickness absence among nurses. *Int Arch Occup Environ Health* 2018;91:195–203.
- 30 Notenbomer A, van Rhenen W, Groothoff JW, *et al.* Predicting long-term sickness absence among employees with frequent sickness absence. *Int Arch Occup Environ Health* 2019;92:501–11.
- 31 Bosman LC, Roelen CAM, Twisk JWR, *et al.* Development of prediction models for sick leave due to musculoskeletal disorders. *J Occup Rehabil* 2019;29:617–24.
- 32 Garde AH, Begtrup L, Bjorvatn B, *et al.* How to schedule night shift work in order to reduce health and safety risks. *Scand J Work Environ Health* 2020;46:557–69.
- 33 Rosenström T, Härmä M, Kivimäki M, *et al.* Patterns of working hour characteristics and risk of sickness absence among shift-working hospital employees: a data-mining cohort study. *Scand J Work Environ Health* 2021;47:395–403.
- 34 Larsen AD, Ropponen A, Hansen J, *et al.* Working time characteristics and long-term sickness absence among Danish and Finnish nurses: a register-based study. *Int J Nurs Stud* 2020;112:103639.
- 35 Shiri R, Martimo K-P, Miranda H, *et al.* The effect of workplace intervention on pain and sickness absence caused by upper-extremity musculoskeletal disorders. *Scand J Work Environ Health* 2011;37:120–8.
- 36 Roos E, Lallukka T, Lahelma E, *et al.* The joint associations of smoking and obesity with subsequent short and long sickness absence: a five year follow-up study with register-linkage. *BMC Public Health* 2017;17.
- 37 WHO European health information at your fingertips. Available: https://gateway.euro.who.int/en/indicators/mn_survey_19-cut-off-for-bmi-according-to-who-standards/ [Accessed 20 Sep 2024].
- 38 Hinderliter AL, Sherwood A, Craighead LW, *et al.* The long-term effects of lifestyle change on blood pressure: one-year follow-up of the ENCORE study. *Am J Hypertens* 2014;27:734–41.
- 39 Grimani A, Aboagye E, Kwak L. The effectiveness of workplace nutrition and physical activity interventions in improving productivity, work performance and workability: a systematic review. *BMC Public Health* 2019;19.
- 40 Cancelliere C, Cassidy JD, Ammendolia C, *et al.* Are workplace health promotion programs effective at improving presenteeism in workers? A systematic review and best evidence synthesis of the literature. *BMC Public Health* 2011;11.
- 41 Marin-Farrona M, Wipfli B, Thosar SS, *et al.* Effectiveness of worksite wellness programs based on physical activity to improve workers' health and productivity: a systematic review. *Syst Rev* 2023;12:87.