Using machine learning methods and Annual Population Survey data to predict job loss amongst workers in the early stages of sickness absence

Michael Oldridge, Daniella Murphy, Mariam Ahmed, and Kisshor Samyrao.

Non-technical summary

The number of working-age people who are economically inactive due to ill-health has been increasing in the UK since early 2019. This is partly because once people move out of work due to ill-health, they tend to remain inactive for a long time. Early workplace-based support for workers with health conditions (such as occupational health) can be effective at preventing health-related job losses, particularly when initiated at an early stage. However, evaluations of publicly funded or commissioned occupational health pilots have found that they are often poorly targeted, with people who are not at risk of job loss being referred. This reduces their cost-effectiveness, and prevents resources from being allocated to those most in need.

To improve the quality of referrals, we have built and compared four models that estimate the percentage probability that a worker in the early stages of sickness absence (0-5 weeks) will move out of work 12 months later. This uses data on a wide range of factors relating to socio-demographics, work, health, perceptions about the impact of health conditions on day-to-day activities, family, housing, and education.

Our strongest model is reasonably accurate at predicting job loss, but not highly accurate. This means that the model can correctly identify people who will move out of work (the ‘true positive rate’), but it will also incorrectly predict job loss amongst people who will not move out of work (the ‘false positive rate’). This depends on the probability threshold that is chosen to determine if someone will move out of work or not, and users of the model would need to set the threshold depending on the relative benefits and costs of true positives and false positives in their policy context. However, to illustrate the predictive ability of the strongest model, an 8% decision threshold correctly predicts job loss for 77% of the people who do move out of work and incorrectly predicts job loss amongst 44% of the people who actually retain work.

We also identified important factors for predicting job loss. The factors identified across both of our two strongest models are: age, number of months continuously employed, number of health conditions, self-reported health rating, mental health disability, disability due to back/neck pain, working part-time, a perception that a condition limits day-to-day activities “a lot”, and not having a mortgage.

Prediction models such as this could be trialled in work and health support services. If they are found to be more effective than human judgement at predicting job loss amongst early absentees, using them to generate referrals could improve the effectiveness and cost-effectiveness of the service. Alternatively, referrers could use the estimated probabilities from a prediction model to supplement their own judgement about whether someone is at risk of job loss and would benefit from the service. Even if a prediction model cannot be used, referrers could consider the specific risk factors identified by this study as most important, to support their own judgement.

1. Introduction

The number of working-age people who are economically inactive due to ill-health has been increasing in the UK from 2.3 million people in Aug-Oct 2019 to 3.0 million in Aug-Oct 2024 (ONS, 2024). Having people out of the workforce due to ill-health has significant social and economic costs. Individuals out of work have lower incomes, are at risk of social exclusion, and can face a further deterioration in health and wellbeing. Each year an individual is out of work and on disability benefits instead of working full-time costs £15,000 in lost tax receipts, disability benefits costs, and NHS costs (DWP, 2021). This hampers economic growth. As a result, the UK Government is interested in developing policies that can reduce health-related job loss and keep people in the workforce (DWP/HMT/DfE, 2024) .

Prolonged absences and movements out of work can lead to a further deterioration in health and increasingly complex barriers to returning (Waddell & Burton, 2006). As a result, the longer a sickness absence (SA) lasts, the less likely an individual ever returns to work (DWP, 2019). Once someone moves out of work due to ill-health, they tend to remain economically inactive for a long period. However, certain types of early, in-work support for employees with health conditions, such as Occupational Health (OH) services, can prevent employee ill-health from leading to long-term SAs, job losses, and long-term inactivity (Waddell, Burton, & Nicholas, 2008) (Wynne-Jones, et al., 2018) (Cancelliere, et al., 2016).

These services are more effective when delivered to those genuinely at risk of job loss and when initiated at an early point in their sickness journey (Waddell, Burton, & Nicholas, 2008). For example, Cancelliere et al. (2016) found they are more effective when initiated in the first six weeks of an absence. However, there is no established screening method for identifying workers who are at risk of job loss in the early stages of their SA. Length of SA is often found to be most important predictor of job loss, with people who have been off sick for longer more likely to leave work, but this does not allow us to predict job loss early in SA. As a result, publicly funded OH pilots tend to either restrict eligibility to workers whose SA has lasted a minimum number of weeks, or alternatively allow referrers (most often GPs and employers) to use their own judgment about who should be referred. This means interventions can be too late to prevent conditions deteriorating, and/or referrals may be unsuitable. For example, in the Study of Work and Pain (SWAP) trial, many participants referred by GPs felt their situation was not sufficiently serious for them to require the support, because it was a self-limiting condition, it wasn’t affecting their ability to work, or they already had support in place through their employers (Sanders, Wynne-Jones, Nio Ong, Artus, & Foster, 2019). Similar issues about suitability of referrals were also reported in other public OH pilots (Demou, Hanson, Bakhshi, Kennedy, & Macdonald, 2018) (Bebb & Heledd, 2019) (Batty, et al., 2021). Unsuitable referrals limit the effectiveness and cost-effectiveness of the service, and reduce service availability for those who need the support.

In this report, machine learning methods and Annual Population Survey (APS) data have been used to build a model that predicts, for workers on 0-5 weeks of SA, who might move out of work. Prediction models such as this could be piloted in both public and private work and health services as a means of generating suitable referrals. The efficacy of models at predicting job loss could be trialled against human referrers such as GPs or employers. If superior, it would improve the effectiveness and cost-effectiveness these services by ensuring resources are allocated to those most in need. This would be valuable for stakeholders including OH providers and the UK government in the design of such services.

2. Literature Review

Existing international literature has identified common risk factors for job loss amongst workers with health conditions, and related outcomes such as long-term or repeated SA. A small number of studies have attempted to combine these into screening models that predict outcomes, but none try to predict job loss amongst workers on short-term SA with a variety of conditions and in a variety of industries.

2a. Risk factors

Studies that have explored the risk factors for job loss or SA generally use data from either retrospective questionnaires such as existing OH records (Wilford, et al., 2008), employee data (Notenbomer, van Rhenen, Groothoff, & Roelen, 2019) or APS data (DWP, 2019); or alternatively by collecting new data through prospective cohort studies, for example with primary care patients (Halldén & Linton, 1998), with employees visiting OH services (Bergström, Hagberg, Busch, Jensen, & Björklund, 2014), with disability benefit recipients (Weerdesteijn, et al., 2020), or with employees working for one large employer (Virtanen, et al., 2006).

In terms of methodology, the DWP publication (2019) reported on differences in job retention rates by various demographic and work factors without any modelling. Halldén & Linton (1998) used discriminatory analysis models to identify the most important risk factors. All other reviewed studies used logistic regression to estimate the relationship between different factors and the probability of the outcome. There are lots of potential risk factors and many are closely related to each other. Therefore, many published studies first reduced dimensionality (the number of independent variables or factors) and collinearity (correlated independent variables or factors) amongst the predictors. They did this by removing those that were not significant in univariate models, removing those that were highly correlated to each other, and/or through stepwise model selection. They then used smaller sets of the most significant variables in final multivariate models, to identify the most important factors.

Systematic reviews have summarised findings from these studies. For example, Blank et al. (2008) reviewed 15 studies to draw conclusions about common risk factors for job loss amongst people with mental health (MH) conditions on less than six months of SA, and Cancelliere et al. (2016) systematically reviewed other systematic reviews to draw conclusions about job loss risk factors for people on all types of SA.

However, there are limitations in this literature. First, most studies used very specific population groups, with uncertain generalisability to the wider population of UK workers with health conditions. Blank et al. (2008) and Cancelliere, et al. (2016) found that whilst there was an abundance of studies focusing on musculoskeletal (MSK) conditions, the evidence base for people with MH conditions was limited. For example, Halldén & Linton (1998) included only Swedish primary care patients with neck/back pain. In a similar study, Bergström et al. (2014) included only employees visiting an OH service due to neck/back pain, and all of the sample worked in either a paper mill, a truck manufacturer, or a steelworks. Weerdesteijn, et al. (2020) only covered workers with 'subjective health complaints'. The sample in Wilford et al. (2008) covered all health conditions, but only for Scottish public sector employees who had been referred for an OH assessment. Similarly, the participants in the study by Virtanen et al. (2006) only worked in Finnish municipals and hospitals. Only one of the reviewed studies covered a wide variety of industries and health conditions, but this was for Dutch workers only (Notenbomer, van Rhenen, Groothoff, & Roelen, 2019). Aside from a DWP statistical publication which had no modelling (DWP, 2019), there is a lack of literature covering a wide group of UK workers. A second, related issue is potential bias in studies where the samples are sourced from lists of employees who are referred to OH assessments (Bergström, Hagberg, Busch, Jensen, & Björklund, 2014) (Wilford, et al., 2008). Being referred for support means they have already been selected as at risk in some way. On the other hand, the support they receive might alter their outcomes by effectively supporting them to return to work.

Despite these limitations, many risk factors are commonly identified across multiple studies with different methods, population groups, and countries (see table 1). This provides reassurance that they are generalisable to a wide variety of characteristics and circumstances.

Table 1: job loss risk factors commonly identified across multiple studies

Category	**Risk factor**
Socio-demographic	Older age Female Lower educational attainment Being divorced/ widowed/ single.
Work factors	Lower job grades/ socioeconomic status High job stressors Physical work demands
Health factors	Previous sickness absence Mental health conditions Higher severity of symptoms
Psychological factors	Patient perceptions of how conditions affect their ability to perform day-to-day activities Perceptions of their ability to work, and perceptions of their likelihood of return to work

Some authors have hypothesised that that clinical factors, such as health condition and severity of symptoms, are relatively more important for predicting job loss in the earlier stages of a SA and less important in the later stages, whereas socio-demographic and psychological factors are relatively less important for predicting job loss in the earlier stages of SA and more important in the later stages (Steenstra, et al., 2017) (Weerdesteijn, et al., 2020). However, a systematic review found that whilst there was some evidence to support this, the evidence base was inconclusive, largely because socio-demographic and work environment factors are too rarely included in studies on the later stages (Steenstra, et al., 2017).

2b. Prediction models

A much smaller number of studies have tried to combine risk factors into screening models that predict job loss or SA. One example is the Örebro MSK Pain Screening Questionnaire (ÖMPSQ), which collected data on 21 psychosocial factors from patients with MSK pain (Halldén & Linton, 1998). Halldén & Linton identified the five most important factors for predicting SA using univariate and multivariate discriminatory analysis models on a sample of 137 Swedish primary care patients with back/neck pain. These variables included a patient's belief that they should not work with the current pain, perceived chance of recovery, perceived ability to do light work, stress levels, and previous SA. A discriminatory analysis prediction model based on these five variables was then evaluated on the same sample. This appeared accurate at predicting accumulated SA, with 73% accuracy, a 75% true negative rate (TNR), a 77% true positive rate (TPR) for predicting 1-30 days of SA and 61% TPR for predicting 31+ days of SA. No score was provided for the Area Under the Curve (AUC) of the Receiver Operating Curve (ROC). They also evaluated a screening model with all 21 variables, where each was scored on a ten-point scale and weighted equally (i.e. no modelling). The total possible score ranged from zero (high risk) to 210 (low risk). This also appeared accurate. A cut-off score of 105 led to a 75% TNR, an 86% TPR for 1-30 days of SA, and an 88% TPR for 31+ days.

The ÖMPSQ has subsequently been used many times for predicting various outcomes on different populations. A systematic review compared prediction scores from 14 studies which used the ÖMPSQ, as well as an almost identical questionnaire focused specifically on neck/back pain (Sattelmayer, Lorenz, Röder, & Hilfiker, 2012). Out of the four studies that focused on job loss/return to work as the outcome, the TPRs were 1.00, 0.79, 0.45, and 0.19, and the corresponding TNRs were 0.62, 0.59, 0.94, and 0.94, respectively. This heterogeneity was partly due to different sample populations and partly due to the use of different cut-off scores, which varied from 105-147. Unfortunately, the review did not compare a threshold-invariant measure such as AUC of the ROC curve. Later, in 2014, a study again tested the ÖMPSQ on 195 employees who worked for 4 large male-dominated businesses in 2000-2001 and visited an OH service due to neck/back pain (Bergström, Hagberg, Busch, Jensen, & Björklund, 2014). They found it was highly accurate at predicting 'disability pension', with an AUC of 0.93.

Wilford et al. (2008) built a job loss prediction model using logistic regression and five variables that they had identified as most important through univariate models and stepwise model selection. These included a patient's perceptions of their chances of return to work, their perceptions of their ability to work in six months' time, previous SA, age, and whether they were waiting for treatment. Using the same sample for evaluating predictive performance, they found it was highly accurate, with an AUC score of 0.90.

More recently, Notembomer et al. (2019) built similar prediction models using two sets of variables they had identified as important through logistic regression and stepwise model selection. Both models included age, gender, education, marital status, and prior long-term SA. Model 1 additionally included work pace, role clarity, and learning opportunities, whereas model 2 additionally included burnout and work engagement. These two sets of variables were used in two separate logistic regression models to predict long-term SA (lasting at least six weeks), based on a sample of 3,563 Dutch workers who had at least 3 SA spells in the previous 12 months. To reduce overfitting, these were validated in 250 bootstrapped samples. Both the first and second models had some predictive ability, with bootstrap-adjusted AUC scores of 0.615 and 0.616 respectively, but this was not considered strong enough to be useful for practice.

The limitations discussed in section 2a in relation to risk factors also apply to many of the prediction models. Firstly, most of the models are only built and tested on very specific population groups. The Notembomer et al. (2019) study was the only one which covered a wide range of health conditions and industries, but participants had to have had at least 3 SA spells in the previous year. It was also not based on UK workers. Secondly, the studies using samples where someone has already been referred to OH may have biased results.

An additional limitation of the logistic regression prediction models is the risk of overfitting to the training data, which causes low training bias but high variance. Unfortunately, this is not testable because all reviewed models evaluated their accuracy only on the same sample they were built on, meaning we cannot evaluate out-of-sample performance. As variance is higher in models with a high number of features relative to the number of observations, overfitting is particular risk in the models based on small sample sizes of below 200 (Halldén & Linton, 1998) (Sattelmayer, Lorenz, Röder, & Hilfiker, 2012) (Bergström, Hagberg, Busch, Jensen, & Björklund, 2014). The variance may be lower where studies have first reduced the number of predictors through stepwise model selection, but this does not necessarily eliminate overfitting. To further reduce variance, these studies would benefit from machine learning methods such as penalised regression or tree-based methods with cross-validated hyperparameters. The only study that did attempt more advanced methods to reduce variance was Notembomer et al. (2019), by validating their results with 250 bootstraps. However, bootstraps only work where forecasts from each bootstrapped sample are independent of each other, which is unlikely if the bootstrapped samples use some of the same features. Regardless, Notembomer et al. (2019) had a much larger sample size than most of the other studies and so variance is likely to be lower in this study. The studies using the full OREMBRO questionnaire were not models, so those scores were not at risk of overfitting. However, they all had low sample sizes of below 200, so their prediction scores are still likely to have high variance.

Finally, a notable gap in previous prediction models is that none have attempted to predict job loss specifically amongst people on short-term SA (e.g. 0-6 weeks). This is a significant gap since OH interventions are more effective when initiated at an earlier stage (Waddell, Burton, & Nicholas, 2008) (Cancelliere, et al., 2016).

3. Data

This study used fives datasets from the two-year longitudinal APS (2015-16, 2016-17, 2017-18, 2018-19, and 2019-20). This is a survey of working-age people in the UK with data on a range of demographic, social, and economic variables, collected across two waves, twelve months apart. Some of the required APS variables were only available in separate cross-sectional APS datasets, so a unique person identifier was used to merge these from the cross-sectional APS onto the longitudinal data. The five longitudinal datasets were then appended together into a pooled cross-sectional dataset with a row per unique person, variables for that person at one moment in time (wave one), and an additional variable that describes employment status 12 months later (wave two). The data was filtered to include only people who, at the time of wave one, were under retirement age, were employed or self-employed, and were on sick leave for up to five weeks. Any observations with any missing values that could not be explained were removed, other than the gross weekly wage variable which was replaced with the median wage because removing the observations with missing values would remove all self-employed people from the sample. This resulted in 1,541 observations, 61 variables from wave one, and one variable from wave two. The full variable list is given in Annex A. Many of these variables were categoric, so after the categories were converted into separate dummy variables, the total number of features increased to 300.

4. Method

The primary goal of the study was to predict how many of our sample had left employment by wave two (12 months later). The secondary goal was to identify important risk factors in wave one.

The 61 variables from wave one (see annex A) can be grouped into six broad categories: socio-demographic (7 variables), work (15), health and psychological (31), family and housing (6), education (1 variable), and month of the year the individual was interviewed for the APS (1 variable). The data included most of the risk factors identified in existing literature, other than job stressors, work demands, previous SA, or an individual's perceived likelihood of their return to work. The latter two are particularly significant omissions as they were two of the most commonly identified risk factors in existing literature.

Given the dataset had a large number of features relative to the number of observations, there was a risk our models could overfit to the training data, resulting in low training bias but high variance. This means they could perform well in-sample but poorly out-of-sample. We therefore used Python to apply machine learning methods which can reduce the variance and the risk of overfitting. We used 75% of the sample as training data, and retained 25% as testing data (stratified because the data is unbalanced). In order to identify the model with the best out-of-sample predictive strength, we trained four different models on the training data and compared their predictive strength on the testing data using the AUC. The four models included multivariate logistic regression with a Ridge penalty, multivariate logistic regression with a LASSO penalty, a pruned Classification Tree, and a Random Forest. For the two models with the best testing AUC, we then considered how the accuracy, TPR, false positive rate (FPR), and precision would vary based on using different probability thresholds for determining classes. This is because best threshold would depend upon the specific intervention that the model is being used for, and a thorough assessment of its relative costs and benefits. For more information on and justification of the methodological approach, see annex B.

5. Modelling results

5a. Predictive ability

As shown by table 2, all models overfitted to the training data in terms of ROC AUC scores, as they all performed worse on the testing data. The model with the best AUC scores on the testing data was the Random Forest (0.735), followed by the LASSO logistic regression (0.680). These models appear to have some predictive ability, although they might not be considered highly accurate. The Ridge logistic regression had a weaker AUC score (0.647), and the Classification Tree was little better than a random prediction (0.510).

Table 2: Training and testing accuracy and ROC AUC scores

Model	Training accuracy	Testing accuracy	Training ROC AUC	Testing ROC AUC
Ridge	0.926	0.917	0.898	0.647
LASSO	0.919	0.920	0.722	0.680
Classification Tree	0.919	0.920	0.845	0.510
Random Forest	0.919	0.920	0.879	0.735

The AUC score of our Random Forest model was lower than the scores given in the Wilford et al. (0.900) and Bergström, et al. (0.930) studies. However, these studies are not directly comparable as they were tested on narrower population groups and the Bergström, et al. study predicted SA rather than job loss. Both scores were also uncertain due to having high variance and the risk of bias (discussed in section 2b). On the other hand, our Random Forest has a higher AUC score than the Notenbomer, et al. models (0.615/0.616), which also covered a wider range of workers and attempted to reduce overfitting (Notenbomer, van Rhenen, Groothoff, & Roelen, 2019). However, that model predicted long-term SA rather than job loss so it is again not directly comparable.

Based on a 50% decision threshold for determining classes, all four of our models had a very high accuracy score on the testing data of around 92%. However, as explained in annex B, accuracy at a 50% decision threshold is not a good metric to evaluate models that are based on imbalanced data such our ours. For example, the LASSO, Classification Tree and the Random Forest all achieve this high accuracy just by predicting every case in the testing sample will remain in work (see confusion matrices in annex D). This is not useful, because it would not fulfil the aim of this study which is to identify people who will fall out of work.

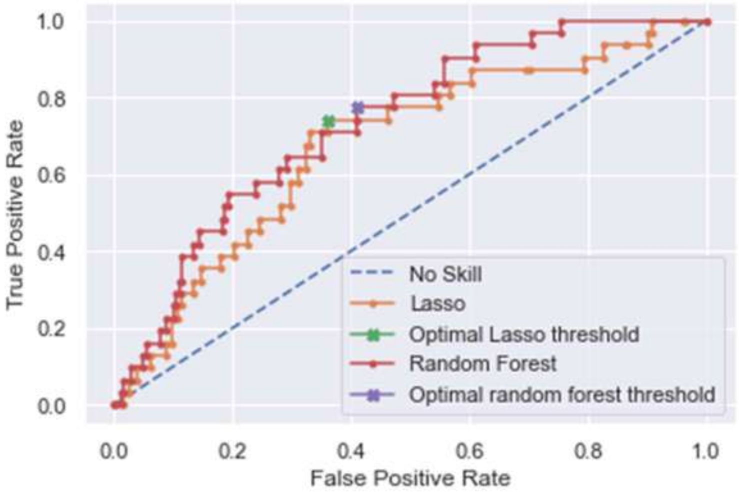
In figure 1, we present the ROC curves for LASSO and Random Forest as the two best performing models. This plots the TPRs and FPRs for all decision thresholds. On both curves, the 50% decision threshold is the point in the far bottom left-hand corner, where the TPR and FPR are both 0%. The points closest to the top-left hand corner represents the optimal thresholds if we valued the TPR and specifity equally, and are calculated using the Youden J statistic. This point is highlighted in the ROC curves in figure 1, the prediction scores are presented in table 3, and the confusion matrices are given in annex D. For LASSO, this optimal point has a threshold of 8%, a TPR of 74%, a FPR of 36%, an accuracy of 65%, and a precision of 15%. For the Random Forest, this point has a threshold of 8%, a TPR of 77%, a FPR of 41%, an accuracy of 60%, and precision of 14%.

Table 3: Testing prediction scores at different decision thresholds for LASSO and Random Forest

Model	Decision threshold cut-off	TPR	FPR	accuracy	precision
LASSO	50%	0%	0%	92%	Na

RF	50%	0%	0%	92%	Na
LASSO	8%	74%	36%	65%	15%
RF	8%	77%	41%	60%	14%

Figure 1: ROC curves for LASSO and Random Forest



5b. Important variables

For the LASSO logstic regression, see Annex E for the the features with coefficients that have not been shrunk to exactly zero by the LASSO penalty. Annex E also shows for the variables in the Random forest with a features importance score of over 0.01. Reassuringly, many variables are identified in both methods. These include:

- Age
- Number of months continuously employed
- Number of health conditions
- Self-reported health rating
- Being disabled due to a MH condition
- Being disabled due to a back or neck problem
- Working hours (full-time or part-time)
- Perception that a health condition limits day to day activities “a lot”
- Having a mortgage

The LASSO model additionally selected disability due to hearing loss, working in the energy and water industry, being divorced, having a degree, and having no qualification as important variables. Many of these variables are consistent with the risk factors identified in existing literature. However, we have also identified new variables that have not commonly been identified in previous studies. This includes length of time continuously employed, number of health conditions, having a disability due to back or neck pain, having a disability due to hearing loss, working hours, working in the energy and water sector, and having a mortgage. There could be several reasons these factors have not been identified in previous studies, including that they have not been tested for before, they can only be picked up using machine learning methods, they do not reflect some of the narrow populations used in previous studies, and/or they are only relevant for workers who are on up to 5 weeks of SA. It is also important to note that the fact we have tested a variable but not identified it as important does not mean it is not a job loss risk factor.

6. Conclusions and implications

Our best model, the Random Forest, has decent predictive ability with an ROC AUC score of 0.725. The LASSO and Ridge logistic regressions also had some predictive ability but they were not as strong. At a 50% decision threshold, the Random Forest has high accuracy of 92% but this is not useful for practice because it just predicts everyone in the sample will stay in work. Alternatively, if we value the TRP and specifity equally, then the optimal decision threshold for the Random Forest is 8%. On our testing data, this threshold led to a TPR of 77%, a FPR of 41%, an accuracy of 60%, and a precision of 14%.

By cross-checking the findings from the two strongest models, the Random Forest and LASSO, we identified variables that were important for predicting job loss. Many of these are consistent with those identified in previous studies, but we also found new variables that have not previously been identified, such as length of time continuously employed, number of health conditions, being disabled due to back or neck pain, working hours, and having a mortgage.

This fills a number of crucial gaps in the existing literature. Firstly, as far as we could find, this is the first model that predicts job loss for workers in the first 5 weeks of SA. Secondly, by using APS data, our findings are likely to be generalisable to a wider group of UK workers than previous studies. Thirdly, as far as we could find, this is the first study that predicts job loss amongst absentees by using machine learning methods that reduce overfitting such as penalised logistic regression, tree-based methods, and cross-validation. It is also the first study that evaluates model performance on different data to the data it was built on, so the AUC scores provided are more likely to reflect true out-of-sample performance than the AUC scores given in previous studies. Fourthly, this model included and identified some risk factors that have not been tested before. On the other hand, a significant weakness of this model was that data was not available on two factors which are commonly identified in existing literature: previous SA spells and a patient's own perceptions about whether they will return to work. It is possible that the prediction strength of our models would be even stronger if we could include these factors.

Using the longitudinal APS data for this project had many advantages. Firstly, the survey covers the whole UK working age population, so the results are likely to be generalisable to a wider group of workers than previous UK studies. Secondly, the APS is a large-scale survey, so we have a larger sample size than many previous studies. Thirdly, the APS collects data on a wide set of variables, which enabled us to include many of the factors identified in existing literature as well as additional factors that have not been tested before. There are also downsides to using the APS data. Firstly, it does not include all of the risk factors that have been identified in existing literature. Secondly, we could not pick up cases where someone moved out of work and returned to new work all before the second wave. Thirdly, some of the people will leave work for non-health reasons (e.g. for early retirement or to become a student), in which case providing them with work and health interventions may not prevent the job loss. However, given all of our sample are on sick leave at wave one, they must all have health conditions that affect their work, so it is reasonable to assume that most of the subsequent job losses are at least somewhat related to health.

Further research could build on this study by developing new machine learning models that can be built into referral systems for expert work and health services. Their predictive ability could then be tested against human judgement of existing referrers like GPs. If they are more accurate , they could improve the effectiveness and cost-effectiveness of these services by ensuring the support is allocated to people who are most in need. Alternatively, humans could use the estimated probabilities from the models to inform their own judgement, instead of replacing it. Even where prediction models cannot be used, referrers could at least consider the specific risk factors that have been identified in this study, to support their own judgement about whether someone is at risk of job loss and would benefit from the service.

Annexes

Annex A: Full variable list from APS

		Description	Type	Category
1	AGE1	Age	Numeric	Demographic
2	CIGSMK16	Smoker status	Categoric	Health
3	CRY12	Country of birth	Categoric	Demographic
4	DISEA1	Disability status	Categoric	Health
5	EMPMON1	No. months continuously employed	Numeric	Work
6	ETHUKEUL1	Ethnicity	Categoric	Demographic
7	FDPCH161	No. dep children under 16	Numeric	Family and housing
8	FTPT1	Works full-time/ part-time	Binary	Work
9	GRSSWK1	Gross weekly earnings	Numeric	Work
10	hc_armhand	Disability due to arms/hands problems	Binary	Health
11	hc_autism	Disability due to autism	Binary	Health
12	hc_backneck	Disability due to back/neck problems	Binary	Health
13	hc_breath	Disability relating to breathing, asthma, or lungs	Binary	Health
14	hc_diabetes	Disability due to diabetes	Binary	Health
15	hc_dig_liv_kid	Disability due to stomach, liver kidney or digestive problems	Binary	Health
16	hc_epilepsy	Disability due to epilepsy	Binary	Health
17	hc_hearing	Disability due to hearing problems	Binary	Health
18	hc_heartblood	Disability due to heart, blood pressure or blood circulation	Binary	Health
19	hc_learning	Disability due to learning difficulties	Binary	Health
20	hc_legsfeet	Disability due to legs/feet problems	Binary	Health
21	hc_mh	Disability due to depression, bad nerves or anxiety	Binary	Health
22	hc_mh2	Disability due to mental illness or nervous disorders	Binary	Health
23	hc_other	Disability due to other health problems	Binary	Health
24	hc_progressive	Disability due to progressive illness not included elsewhere	Binary	Health
25	hc_sight	Disability due to sight problems	Binary	Health
26	hc_skin	Disability due to Severe disfigurement, skin conditions, allergies	Binary	Health
27	hc_speech	Disability due to speech	Binary	Health
28	HEALTH	Main health condition type	Categoric	Health
29	HIQUL15D1	Highest qualification	Categoric	Education
30	HOME1	Works from home status	Categoric	Work
31	illcause	Condition that caused SA	Categoric	Health
32	INDE07M1	Industry 9 groups	Categoric	Work
33	JBTP101	Type of temporary job	Categoric	Work
34	JOBTYP1	Job permanent or temporary	Categoric	Work
35	LIMACT1	Does long-term health condition limit day to day activities	Categoric	Health
36	LIMITA	Does health condition affect amount of paid work can do	Categoric	Health
37	LIMITK	Does health condition affect kind of paid work can do	Categoric	Health
38	LLORD1	If renting, landlord type	Categoric	Family and housing
39	LNGLST1	Do you have a long-term health condition	Categoric	Health
40	lost_job	lost job 12m later	Categoric	Outcome
41	MANAGER1	Managerial status	Categoric	Work
42	MARSTA1	Marital status	Categoric	Family and housing
43	MPNR02	No. employees at work	Categoric	Work
44	Numberhcs	No. health conditions	Numeric	Health
45	NUTS132	NUTS132 area	Categoric	Demographic
46	PHEAL_LIM	Past long-term health condition that limited activity	Binary	Health
47	PHEAL_NOLIM	Past long-term health condition that did not limit activity	Binary	Health
48	PUBLICR1	Public or private organisation	Categoric	Work
49	QHEALTH1	Health rating 1-5 (1; very good, 5: very bad)	Ordinal	Health
50	REDACT	Length of time activities been reduced	Ordinal	Health
51	REFWKM	Month of APS interview	Categoric	Interview month
52	REGWKR1	Region of place of work	Categoric	Work
53	RELHFU1	Relationship to head of household 3 groups	Categoric	Family and housing
54	RELIG11	Religion	Categoric	Demographic
55	RU11IND1	Rural/urban classification	Categoric	Demographic
56	SC10MMJ1	Occupation 9 groups	Categoric	Work
57	SECTRO031	Type of non-private business	Categoric	Work
58	SEX	Sex	Categoric	Demographic
59	SOLOR1	If self-employed, do you have employees	Categoric	Work
60	STATR1	Main job employment type	Categoric	Work
61	TEN11	Housing type	Categoric	Family and housing
62	TIED	If renter, is accommodation tied to job or not	Categoric	Family and housing

Annex B: Further detail on methods

Out-of-sample predictive ability is the primary goal, but we also aim to identify the most important variables for prediction in the models with the best predictive strength. This is for two reasons. Firstly, not all of the data we have used will be available in the IT systems of people who could refer to a public OH service, such as GPs or employers, and it would not be feasible for them to collect all of it in order to make these predictions. It could however be feasible for them to collect data on a smaller set of variables. Secondly, even if the prediction model is not used, identifying the most important variables will contribute to the existing literature about risk factors for job loss. This can be useful for practitioners to consider when making their own judgements about whether someone is at risk, and useful for researchers in attempting to build more accurate prediction models in the future. This will be particularly useful given we have considered factors not tested in previous studies.

The four models we have chosen are multivariate logistic regression with a Ridge penalty, multivariate logistic regression with a LASSO penalty, a pruned Classification Tree, and a Random Forest. All of these methods can be used for classification problems, all allow us to utilise machine learning methods, and all enable some level of interpretation. For example, the Ridge and LASSO methods both introduce a penalty into the Gini index loss function, which penalises having a high number of features relative to the number of observations. This shrinks the coefficients, which reduces model complexity and reduces the risk of overfitting. To do this, we first scale the features so they are centred around 0 and the standard deviation is equal to 1. We then use k-fold cross-validation (k=5) to determine the Ridge and LASSO penalties that lead to the strongest models in terms of the AUC. As in standard logistic regression, both Ridge and LASSO provide log odds coefficients that allow us to identify the most important variables for the model. With Ridge, we can identify the variables with the largest coefficients. The LASSO penalty can shrink the coefficients of less important variables to zero, so in that model we can see which variables have a non-zero coefficient after the LASSO penalty has been applied.

We also use tree-based methods because they are well suited to dealing with challenges associated with having lots of predictors, such as multicollinearity and interactions between predictors. Incorporating interactions between predictors may be particularly important in this context, as it’s possible that there are different job loss risk factors for people with different health conditions, for people with different characteristics, or for different industries. To reduce overfitting in the Classification Tree, we use cost-complexity pruning, and apply this by cross-validating (k=5) the maximum depth and the minimum sample per leaf. However, even pruned Classification Trees are prone to overfitting, so we also build and test a Random Forest. Random Forests reduce variance and the risk of overfitting by bootstrapping the training data, using non-overlapping subsets of features within each bootstrap to remove correlation between forecasts, and then aggregating forecasts from the trees built on each bootstrapped sample. We again use k-fold cross-validation (k=5) to identify the maximum depth, minimum samples per leaf, and number of trees that achieve the best AUC score in the Random Forest. To identify the most important variables in the pruned Classification Tree, we can identify which features and splits have been used in the final tree, and we can identify the most important variables for the Random Forest using variables importance.

For the two models with the best testing AUC, we then consider how the accuracy, TPR, false positive rate (FPR), and precision would vary based on using a different probability threshold for determining classes. We do this for two reasons. Firstly, the full dataset is highly imbalanced, meaning the vast majority of cases belong to only one class. In our dataset, 92% of all cases remain in work by wave 2 and only 8% move out of work. This means we could achieve a high accuracy just by predicting everyone stays in work, but this would not be useful because it would not identify anyone who moves out of work, which is the aim of this study. Secondly, true positives and true negatives are not necessarily valued equally in predicting job loss. The economic and social costs of someone falling out of work are potentially very high (e.g. £15k Government fiscal costs per year someone is out of work and on disability benefits rather than working full-time (DWP, 2021)), whereas the cost of an OH assessment can be relatively low (usually £100-£400). Therefore, the cost of not predicting and preventing a job loss (a false negative) would be far higher than the cost of incorrectly predicting a job loss (a false positive), and the benefit of correctly predicting a job loss (a true positive) would be far higher than the benefit of correctly predicting no job loss (a true negative). There could therefore be a strong case for using a lower decision threshold, in order to pick up more true positives (which have a relatively high benefit) at the cost of picking up more false positives (which have a relatively low cost) and having lower precision. Ultimately, the best threshold depends upon the specific intervention being considered and a thorough assessment of the relative costs and benefits, and that is outside of the scope of this paper. However, for illustrative purposes, we identify what the optimal decision thresholds would be if we valued the TPR and specifity equally. This is done using the Youden J statistic, which finds the point that maximises the difference between the TPR and the FPR.

Annex C: Cross validation results

Table 4 shows, for each model, the hyperparameters being cross-validated, the range of values being considered, and the optimal values found through k-fold cross-validation with 5 folds. The optimal Ridge penalty term was 0.01, whereas the optimal LASSO penalty term was 0.05. As the optimal Ridge penalty was the lower bound of the range being considered, it's possible that the model could perform better with an even lower Ridge penalty. Therefore, future research should consider a wider range for the Ridge penalty. The Classification Tree had an optimal max depth of 10 and an optimal minimum sample per leaf of 30. The Random Forest had an optimal max depth of 10, an optimal minimum sample per leaf of 17, and an optimal number of trees of 40.

Table 4: Cross-validation results

Model	Cross-validated parameters	Range of values tested	Optimal values
Ridge	Ridge penalty term	0.01, 0.05, 0.1, 0.5, 1, 5, 10, 100	0.01
LASSO	LASSO penalty term	0.01, 0.05, 0.1, 0.5, 1, 5, 10, 100	0.05
Classification Tree	Max depth Minimum sample per leaf	21 values from 1-100 16 values from 0.5-40	10 30
Random Forest	Max depth Minimum sample per leaf Number of trees	21 values from 1-100 16 values from 0.5-40 12 values from 20-240	10 17 40

Annex D: Confusion matrices for LASSO and Random Forest, at different decision thresholds

LASSO confusion matrix (50% decision threshold)

		Predicted	
		0	**1**
Actual	**0**	355	0
	1	31	0

Random Forest confusion matrix (50% decision threshold)

		Predicted	
		0	**1**
Actual	**0**	355	0
	1	31	0

LASSO confusion matrix (8% decision threshold)

		Predicted	
		0	**1**
Actual	**0**	227	128
	1	8	23

Random Forest confusion matrix (8% decision threshold)

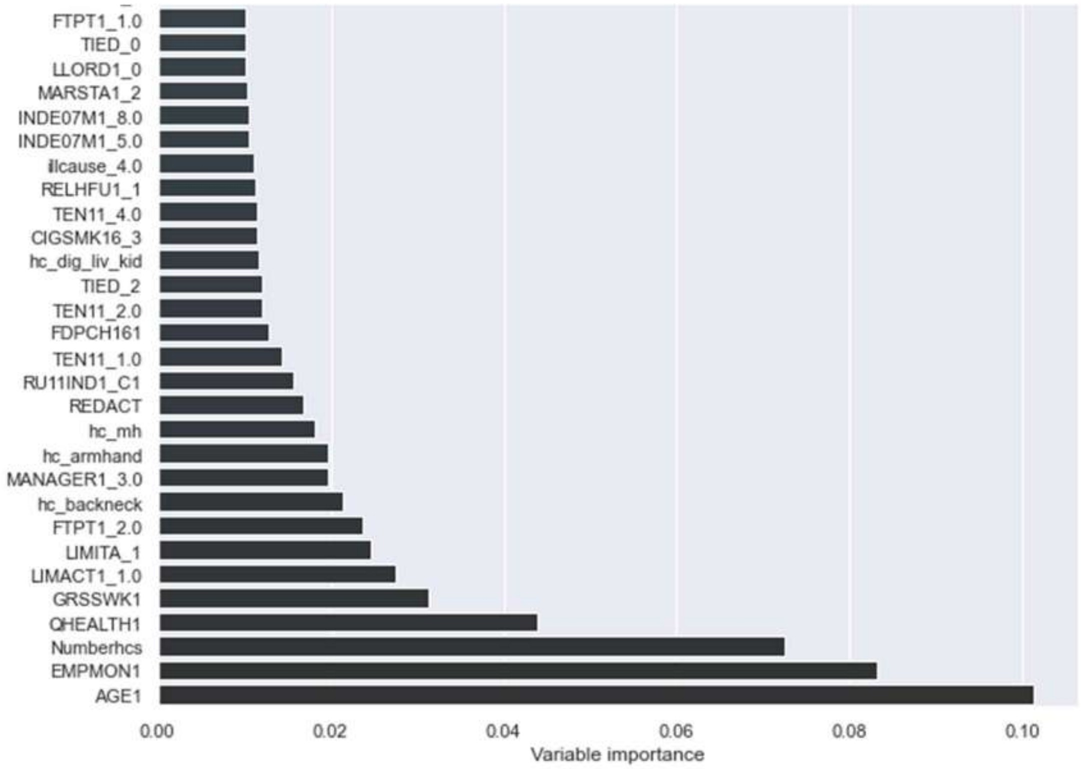
		Predicted	
		0	**1**
Actual	**0**	209	146
	1	7	24

Annex E: Risk factors identified in LASSO and random forest models

Table 5: Variables with non-zero coefficients in LASSO logistic regression model

Variable	Description	Coefficient (log odds)
AGE1	Age	0.0006
EMPMON1	Number of months continuously employed	0.0206
QHEALTH1	Self-reported health rating (1: very good, 5: very bad)	0.0751
hc_backneck	Disabled due to back or neck pain dummy	0.0771
hc_hearing	Disabled due to hearing loss dummy	0.0620
hc_mh	Disabled due to MH condition dummy	0.1402
Numberhcs	Number of health conditions	0.1027
FTPT1_1	Work full-time dummy	-0.0200
FTPT1_2	Work part-time dummy	0.0200
INDE07M1_2.0	Works in Energy and water dummy	0.0180
LIMACT1_1.0	Health condition limits activities 'a lot' dummy	0.0071
MARSTA1_4	Divorced dummy	-0.0044
TEN11_2	Have a mortgage dummy	-0.1384
HIQUL15D1_1.0	Have a degree dummy	-0.0676
HIQUL15D1_6.0	No qualification dummy	0.0029

Figure 4: Variables with features importance scores over 0.01 in the Random Forest



References

Batty, E., Crisp, R., Gilbertson, J., Martin, P., Pardoe, J., Parkes, S., . . . Wilson, I. (2021). *Working well early help annual report 2021.* Project evaluation report, Sheffield Hallam University, Centre for Regional Economic and Social Research. doi:10.7190/cresr.2021.1032611143

Bebb, N., & Heledd, B. (2019). *Evaluation of In-Work Support Operation: final report.* Project evaluation report, Cardiff: Welsh Government. Retrieved from https://gov.wales/evaluation-work-support-operation-final-report

Bergström, G., Hagberg, J., Busch, H., Jensen, I., & Björklund, C. (2014). Prediction of Sickness Absenteeism, Disability Pension and Sickness Presenteeism Among Employees with Back Pain. *Journal of Occupational Rehabilitation, 24*(2), 278-286. doi:10.1007/s10926-013-9454-9

Blank, L., Peters, J., Pickvance, S., Macdonald, E. B., Wilford, J., McMahon, A. D., . . . O'Rourke, A. (2008). A systematic review of the factors which predict return to work for people suffering episodes of poor mental health. *Journal of Occupational Rehabilitation, 18*(1), 27-34. doi:10.1007/s10926-008-9121-8

BoE. (2022). *Monetary Policy Report – November 2022*. Retrieved from https://www.bankofengland.co.uk/monetary-policy-report/2022/november-2022

Cancelliere, C., Donovan, J., Stochkendahl, M. J., Biscardi, M., Ammendolia, C., Myburgh, C., & Cassidy, J. D. (2016). Factors affecting return to work after injury or illness: best evidence synthesis of systematic reviews. *Chiropractic and Manual Therapies, 24*(1). doi:https://doi.org/10.1186/s12998-016-0113-z

Demou, E., Hanson, M., Bakhshi, A., Kennedy, M., & Macdonald, E. B. (2018). Working Health Services Scotland: a 4-year evaluation. *Occupational Medicine, 68*(1), 38-45. doi:https://doi.org/10.1093/occmed/kqx186

DWP. (2019). *Health in the workplace: Patterns of sickness absences, employer support and employment retention*. Retrieved from https://www.gov.uk/government/statistics/health-in-the-workplace-patterns-of-sickness-absence-employer-support-and-employment-ret

DWP. (2021). *Shaping Future Support: The Health and Disability Green Paper evidence pack, July 2021. Chapter 2: Improving Employment Support .* Retrieved from Gov.uk: https://www.gov.uk/government/statistics/shaping-future-support-the-health-and-disability-green-pap

DWP. (2023). *Employment of disabled people 2022.* Retrieved from https://www.gov.uk/government/statistics/the-employment-of-disabled-people-2022/employment-of-disabled-people-2022

DWP/DHSC. (2021). *Heatlh is Everyone's Business: Consultation Response*. Retrieved from https://www.gov.uk/government/consultations/health-is-everyones-business-proposals-to-reduce-ill-health-related-job-loss/outcome/government-response-health-is-everyones-business

DWP/HMT/DfE. (2024). *Get Britain Working White Paper.* Retrieved from Gov.uk: https://www.gov.uk/government/publications/get-britain-working-white-paper/get-britain-working-white-paper

Halldén, K. B., & Linton, S. J. (1998). Can We Screen for Problematic Back Pain? A Screening Questionnaire for Predicting Outcome in Acute and Subacute Back Pain. *The Clinical Journal of Pain, 13*(3), 209-215. doi:https://doi.org/10.1097/00002508-199809000-00007

Macdonald, E. (2022). *Worklessness due to ill health.* Retrieved from Society of Occupational Medicine: https://www.som.org.uk/worklessness-due-ill-health

Notenbomer, A., van Rhenen, W., Groothoff, J. W., & Roelen, C. A. (2019). Predicting long-term sickness absence among employees with frequent sickness absence. *International Archives of Occupational and Environmental Health, 92*, 501-511. doi:https://doi.org/10.1007/s00420-018-1384-6

ONS. (2024). *Dataset INAC01 NSA: Economic inactivity by reason (not seasonally adjusted).* Retrieved from https://www.ons.gov.uk/employmentandlabourmarket/peoplenotinwork/economicinactivity/datasets/economicinactivitybyreasonnotseasonallyadjustedinac01nsa

Sanders, T., Wynne-Jones, G., Nio Ong, B., Artus, M., & Foster, N. (2019). Acceptability of a vocational advice service for patients consulting in primary care with musculoskeletal pain: A qualitative exploration of the experiences of general practitioners, vocational advisers and patients. *Scandinavian Journal of Public Health, 47*(1), 78-85. doi:https://doi.org/10.1177/1403494817723194

Sattelmayer, M., Lorenz, T., Röder, C., & Hilfiker, R. (2012). Predictive value of the Acute Low Back Pain Screening Questionnaire and the Örebro Musculoskeletal Pain Screening Questionnaire for persisting problems. *European Spine Journal, 21*(6), 773-784. doi:DOI:10.1007/s00586-011-1910-7

Steenstra, I. A., Munhall, C., Irvin, E., Oranye, N., Passmore, S., Van Eerd, D., . . . Hogg-Johnson, S. (2017). Systematic Review of Prognostic Factors for Return to Work in Workers with Sub Acute and Chronic Low Back Pain. *Journal of Occupational Rehabilitation, 27*, 369-381. doi:https://doi.org/10.1007/s10926-016-9666-x

The Financial Times. (2023). *Rishi Sunak prepares big push to tackle economic inactivity in the UK*. Retrieved from https://www.ft.com/content/730bbd85-3e15-4e06-8602-c516d4b3e52a

Virtanen, M., Kivimäki, M., Vahtera, J., Elovainio, M., Sund, R., Virtanen, P., & Ferrie, J. (2006). Sickness absence as a risk factor for job termination, unemployment, and disability pension among temporary and permanent employees. *Occupational and Environmental Medicine, 63*, 212-217. Retrieved from https://oem.bmj.com/content/63/3/212

Waddell, G., & Burton, K. A. (2006). *Is work good for your health and wellbeing?* Retrieved from https://www.gov.uk/government/publications/is-work-good-for-your-health-and-well-being

Waddell, G., Burton, K. A., & Nicholas, K. A. (2008). *Vocational Rehabilitation: What Works, for Whom, and When?* Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/209474/hwwb-vocational-rehabilitation.pdf

Weerdesteijn, K. H., Schaafsma, F., Bonefaas-Groenewoud, K., Heymans, M., Van der Beek, A., & Anema, J. (2020). Predicting return to work after long-term sickness absence with subjective health complaints: a prospective cohort study. *BMC Public Health, 20*(1). doi:https://doi.org/10.1186/s12889-020-09203-5

Wilford, J., McMahon, A. D., Peters, J., Pickvance, S., Jackson, A., Blank, L., . . . Macdonald, E. B. (2008). Predicting job loss in those off sick. *Occupational Medicine, 58*(2), Pages 99–106. Retrieved from https://academic.oup.com/occmed/article/58/2/99/1389674

Wynne-Jones, G., Artus, M., Bishop, A., Lawton, S. A., Lewis, M., Jowett, S., . . . Foster, N. E. (2018). *Effectiveness and costs of a vocational advice service to improve work outcomes in patients with musculoskeletal pain in primary care: a cluster randomised trial*. Retrieved from https://pubmed.ncbi.nlm.nih.gov/28976423/