école d'économie d'aix-marseille
aix-marseille school of economics

Working Papers / Documents de travail

The Imaginary Healthy Patient

Amady Seydou Ba
Ewen Gallic
Pierre Michel
Alain Paraponaris

WP 2024 - Nr 35

The Imaginary Healthy Patient

Amady Seydou Ba[1], Ewen Gallic[2], Pierre Michel[2], and Alain Paraponaris[✉2]

[1]Data Engineer/Scientist, Méritis, Paris, France, 36 Av. Pierre 1er de Serbie, 75008 Paris,
[2]Aix Marseille Univ, CNRS, AMSE, Marseille, France

November 2024

Abstract

Anxiety and depression may have serious disabling consequences for health, social, and occupational outcomes for people who are unaware of their actual health status and/or whose mental health symptoms remain undiagnosed by physicians. This article provides a big picture of unrecognised anxiety and depressive troubles revealed by a low score on the Mental Health Inventory-5 (MHI-5) with the help of machine learning methods using the 2012 French National Representative Health and Social Protection Survey (*Enquête Santé et Protection Sociale, ESPS*) matched with yearly healthcare consumption data from the French Sickness Fund. Compared to people with no latent symptoms who did not declare any depression over the last 12 months, those with unrecognised anxiety or depression were found to be older, more deprived, more socially disengaged, at a higher probability of adverse working conditions, and with higher healthcare expenditures backed, to some extent, by chronic conditions other than anxiety or mood disorder.

Keywords

unrecognised mental disorders; mental health inventory-5 (MHI-5); healthcare consumption; workplace outcomes; tree-based methods; SHAP values

JEL Classification Codes

C5; C38; I12

We would like to thank Juliette Spinnato for valuable contributions to the earlier versions of this paper. We also thank Loann Desboulets, Emmanuel Flachaire, Gaëtan Fournier, and Bruno Ventelou for fruitful discussions. This article also benefited from helpful comments from seminar audiences at LEMNA (Nantes) and LEM (Lille) and from participants to the 42nd French Annual Health Economics Conference (Brussels) and the 13th European Health Economics Association Conference (Oslo). At last, we sincerely thank the two anonymous referees whose comments undoubtedly contributed to improve the quality of the paper, as well as the Editor, Tarik Tazdaït, and the two co-editors of the special issue, Arthur Charpentier and Emmanuel Flachaire.

At the time of writing, Amady Seydou Ba was involved in Progexia's NovaCare project, partly funded by Malakoff Humanis, life & health French insurance for employees and individuals.

The project leading to this publication received funding from the French government under the "France 2030" investment plan managed by the French National Research Agency (reference: ANR-17-EURE-0020) and from the Excellence Initiative of Aix-Marseille University—A*MIDEX.

An ebook with replication codes and robustness checks can be accessed here: https://3wen.github.io/imaginary-healthy/.

✉Corresponding author: alain.paraponaris@univ-amu.fr

"He walks, sleeps, eats, and drinks, like other folks, but that does not hinder him from being very ill."

Toinette, Act II, Scene 3, Molière, *The imaginary invalid*, 1673

1 Introduction

Research on mental disorders (anxiety and depression) in the general population has yielded puzzling results. On the one hand, the main part of the literature focusing on this topic states that these mental health troubles are largely unrecognised (under-reported by patients and/or under-diagnosed by health professionals) in the general population (Falagas et al., 2007; Freeling et al., 1985; Higgins, 1994; McQuaid et al., 1999; Sheehan, 2004). On the other hand, less frequent but not less relevant works conclude that mental health problems may be over-diagnosed to some extent (Aragonès et al., 2006; Klinkman et al., 1998). Underdiagnosis and overdiagnosis of anxiety and depressive disorders may lead to detrimental (absence of) care, resulting in a serious mismatch between people who obviously need and those who receive antidepressant drugs and/or anxiolytics (Mojtabai and Olfson, 2011; Druss et al., 2007; Jureidini and Tonkin, 2006; Van der Heyden et al., 2009). Overdiagnosis of anxiety and/or depression may undoubtedly result in unnecessary healthcare expenses (Eveleigh et al., 2014; Luppa et al., 2008), which is beyond the scope of this article. In contrast, unrecognised anxiety-depressive disorders can be fuelled, among other things, by specific social and occupational situations and have strong adverse consequences on health (Falagas et al., 2007), healthcare consumption (Rost et al., 1998; Sheehan, 2004) and occupation (Broadhead et al., 1990; Egede, 2007; Asami et al., 2014; Lim et al., 2000; Simon et al., 2001; Hilton et al., 2010; Stewart et al., 2003). In this respect, the cited literature widely documents the benefits of the diagnosis and treatment of mental health problems, including cost-benefit advantages from a workplace perspective (Berndt et al., 1998; Rizzo et al., 1996; Kessler et al., 1999; Beck et al., 2014). It scrutinises much less the hidden upshots that their non-recognition may create regarding population health, quality of life, social engagement, and participation in the labour force, whereas untreated and treated depression can have serious consequences on absenteeism, presenteeism, and productivity losses (Melchior et al., 2009; Koopmans et al., 2008).

The aim of this paper is to provide an overview of people suffering from unrecognised mental disorders: anxiety and/or depression, *i.e.*, unwell people who are unaware of their illness, unlike Argan, Molière's character, who readily imagines himself ill, and whom we will refer to in the rest of the article as the *imaginary healthy* patients. The analysis relies on a subsample of people aged 18 to 65 years, extracted from the French 2012 Comprehensive Health, Health Care and Insurance Survey (Enquête Santé et Protection Sociale, ESPS), which provides information about health and life outcomes of 23,047 community-dwelling people aged 15 and older. These data are matched with the yearly individual healthcare consumption billings from the National Health Insurance Fund. Descriptive statistics are first used to compare people who do not spontaneously report mental disorders with those with low or high Mental Health Inventory-5 (MHI-5) scores. In addition, a machine learning approach based on various methods (random forest, extreme gradient boosting, support vector machine, penalised logistic regression; see Hastie et al. (2001) for an overview) is used to characterise unrecognised mental disorders and identify discriminant variables associated with them. This approach is preferred over parametric estimation strategies due to the multiple endogeneity and multicollinearity issues present in this cross-sectional design. These methods are well-designed when the outcome is a self-reported health status that can be produced by complex nonlinear relationships and interaction terms (Lamu and Olsen, 2016; Doupe et al., 2019). The contributions and interactions of such factors are assessed using a model explainer method based on Shapley values. The implications of these ignored mental disorders are especially studied regarding concomitant health issues, the existence of potential barriers to accessing care, the level and structure of healthcare consumption, social and family (dis)engagement, working conditions, and quality of life at work. Inconsistencies between the (absence of) mental disorders spontaneously self-reported by individuals and their actual anxiety and depression disorders may signal worrisome situations of mental health needs that are not diagnosed by health professionals and/or recognised by the respondents themselves. These disorders are eventually associated with, but not completely reducible to, problems in access to care. Their solving may help improve the health status and quality of life of people suffering from them.

The remainder of this article is organised as follows. Section 2 describes the population at stake, data collection, and various statistical and machine-learning methods developed in this study. Section 3 provides descriptive statistics, presents the results, and analyses the determinants of self-reported depression, whether valid or contradictory with respect to MHI-5 scores. Section 4 discusses the results. Section 5 concludes.

2 Materials and Methods

2.1 Study Design and Settings

The 2012 Health, Health Care, and Insurance Survey (Enquête sur la Santé et la Protection Sociale, ESPS) is a representative French cross-sectional survey conducted by the Institute for Research and Information in Health Economics (Institut de Recherche et de Documentation en Economie de la Santé, IRDES Paris) comprising 23,047 French residents aged 15 years and older (participation rate: 66%) (see Herr et al. (2017) or Pierre and Jusot (2017) for the presentation of the survey). Respondents were first interviewed either by telephone or face-to-face at home regarding their health status, insurance, and access to healthcare services. Socioeconomic status (SES) was documented using self-administered questionnaires.

The survey data, excluding individual healthcare consumption data, are available on the website of the French National Archive of Data from Official Statistics (Archives de Données Issues de la Statistique Publique, ADISP), and the data including individual healthcare consumption are available on the website of the French Secure Data Access Center (Centre d'Accès Sécurisé aux Données, CASD).

The survey was approved by the French National Commission for Data Protection and Liberties (Commission Nationale Informatique et Libertés, CNIL), which monitors respect for liberties and protection of individual data.

2.2 Data Collection

The data used in this study match the self-reported responses given by the individuals in the ESPS survey to personal billings for healthcare consumption from the French National Health Insurance records (Système National d'Information Interrégimes de l'Assurance Maladie, SNIIR-AM).

Survey data. The 2012 French National Representative Health and Social Protection Survey provides information on respondents' socio-professional characteristics, either at the individual or household level: age, gender, occupation, professional status, social security status and scheme, marital and family status, household income, net income per consumption unit, size of the urban area, zoning in urban areas, and geographical region. Mental health-related quality of life is assessed using the Mental Health Inventory (MHI-5) score, which is an internationally validated instrument issued from the Medical Outcome Study Short-Form 36 (SF-36) (Hoeymans et al., 2004; Rumpf et al., 2001; Thorsen et al., 2013). This score is computed using the responses to five items measuring different aspects of the self-perception of mental burden concerning nervousness, self-motivation, peacefulness, sadness, and happiness in the previous four weeks (Ware et al., 2001).[1] Using a supplementary binary item, respondents were also invited to self-report the presence or absence of a depressive episode in the previous 12 months, as well as any experience with other health-related conditions including asthma, bronchitis, heart attack, artery disease, hypertension, stroke, osteoarthritis, low back pain, neck pain, diabetes, allergy, cirrhosis, and urinary incontinence. Working conditions are assessed using nine items from the Job Content Questionnaire (Karasek, 1985; Karasek et al., 1998). With four response choices (always, often, sometimes, and never), these items capture information about autonomy, skill use, psychological demand, social support, and physical demand. Finally, social conditions are monitored through participation in group projects and the frequency of social interactions.

Healthcare consumption billings from the French health insurance records. These data provide measures of actual ambulatory healthcare consumption (inpatient hospital stays were excluded): number of medical visits, healthcare expenses, reimbursements, co-payments, extra fees, and deductible fees in different medical fields (general practitioners, specialists, pharmacists, physiotherapists, nurses, dentists, equipment, transport, optics, prostheses, and emergencies with or without hospitalisation).

Sunlight data. The literature reports negative correlations between mental health status and the duration of sunlight exposure (Kent et al. 2009; Wang et al. 2019; Kim et al. 2021). To capture this effect in our analysis, we use monthly sunlight records from Météo–France's public data service.[2] The observations are provided by geolocated stations. We retain only stations for which data are consistently available over several years, from 1990 to 2019, to exclude stations that could be ephemeral and thus introduce noise into the data. In total, 57 stations are used. To match monthly meteorological observations given at the scale of weather stations with survey data, where the finest geographical indication is the region and the interview year is 2012, meteorological data must be aggregated to obtain the climate data. Initially, we interpolated sunlight values on a $5° \times 5°$ grid covering metropolitan France using kriging. The resulting data represent the average monthly sunlight exposure values. The grid was created using the GEOFLA database provided by the Institut National de l'Information Géographique et Forestière.[3] Given that the self-assessment question regarding depression concerns the previous year at the time of the interview, we aggregate the interpolated monthly values from the 2012 grid by administrative region.

[1] The five questions the MHI-5 score is based on are: 'How much of the time in the previous four weeks...have you been a very nervous person?', '...have you felt so down in the dumps that nothing could cheer you up?', '...have you felt calm and peaceful?', '...have you felt downhearted and blue?', '...have you been a happy person?', each admitting one of the six answers: 'all of the time', 'most of the time', 'a good bit of the time', 'some of the time', 'a little of the time', 'none of the time'. The value scale applied to the answers to the positively (respectively negatively) formulated questions is: 5, 4, 3, 2, 1 and 0 (respectively 0, 1, 2, 3, 4 and 5).
The individual MHI-5 score is computed by adding the five value scales, the sum of which is multiplied by 4, so that the MHI-5 score ranges from 0 (poorest mental health) to 100 (best mental health).

[2] The weather data are freely accessible at: https://donneespubliques.meteofrance.fr/?fond=produit&id_produit=115&id_rubrique=38.

[3] The GEOFLA database can be accessed freely at: https://www.data.gouv.fr/en/datasets/geofla-r/.

2.3 Statistical Analysis

Descriptive statistics are first computed for the entire study sample, the sample of people who self-report depression and those who do not. Individuals are then discriminated according to their computed MHI-5 score depending on whether it is below or above the first quartile (Q1) of the score distribution. This leads to a 4-class representation of the respondents:

- D_{Q1-}: Individuals who self-report depression, with an MHI-5 score below Q1.
- D_{Q1+}: Individuals who self-report depression, with an MHI-5 score above Q1.
- $\overline{D}_{Q1-}$: Individuals who self-report no depression, with an MHI-5 score below Q1, (*i.e.*, those we call the *imaginary healthy* patients).
- $\overline{D}_{Q1+}$: Individuals who self-report no depression, with an MHI-5 score above Q1.

Of these four classes, the focus is on the last two. Thus, the sample is restricted to those who reported to have not experienced a depressive episode within the previous twelve months. There are two reasons for this choice. First, this study focuses on people with unrecognised mental disorders, including depression. Second, the number of observations in the initial sample of individuals in the first two categories is very low, particularly for those with an MHI-5 score above the first quartile.

The following characteristics are compared between the last two groups: (1) the scores of MHI-5, age, household income, number of medical visits, healthcare expenses and extra fees using mean comparison tests with ANOVA; (2) gender, occupation, professional status, marital status, household type, social security status and scheme, insurance coverage, and geographical region using χ^2-test; and (3) income class, size of urban area and urban area zoning using the Kruskal-Wallis rank sum test.

2.4 Classification

Among those who did not self-report depression, the question of discriminating individuals with a low MHI-5 score ($\overline{D}_{Q1-}$ individuals) and individuals with a high MHI-5 score ($\overline{D}_{Q1+}$ individuals) is addressed as a supervised binary classification problem. We train four types of algorithms to classify individuals into their respective groups. Specifically, we employ three popular machine learning algorithms: two ensemble methods—Random Forests (RF) and Extreme Gradient Boosting (XGBoost)—and a Support Vector Machine (SVM).[4] Additionally, we consider a fourth model, a regularisation method known as Penalised Logistic Regression. The objective is to effectively discriminate between the two groups of individuals (*imaginary healthy* patients and healthy patients) using these widely recognised algorithms.

To train the classifiers, we adopt the following methodology. The sample of individuals reporting no depressive episodes in the past twelve months is split into a training set comprising 80% of the observations and a test set comprising the remaining 20%. The models are trained on the training sample, and their predictive performances are evaluated on the test sample. To select the hyperparameters for the models, we use grid search and repeated k-fold cross-validation (5 folds, 10 repetitions).

One of the main interests of the classification in this study is to predict unrecognised anxiety and/or depressive disorders (individuals in the $\overline{D}_{Q1-}$ group). In doing so, we must pay attention to the measure of sensitivity (or the true-positive rate, when defining the *imaginary healthy* as the positive class). The choice of hyperparameters is made to optimise sensitivity,

[4] More details on the algorithms are provided in Appendix B.1.

in order to maximise the detection of individuals with potential unrecognised mental health disorders.

2.5 Interpretation of Scores

To explain the model predictions, we used Shapley additive explanation (SHAP) values (Lundberg and Lee, 2017). SHAP values are a concept inherited from game theory (Hart, 1989), and were initially used to assess the individual contribution of an agent when considering different agents' coalitions in a cooperative game. In other words, they represent the marginal contribution of an agent where all coalitions have been considered. When applied to predictive machine learning methods such as classification tasks, this method can be used to evaluate the marginal importance of a variable when predicting an outcome. Although tree-based methods already have an embedded method to evaluate the importance of individual variables reflecting their discriminating power to distinguish between several classes of individuals (Breiman, 2001), SHAP values offer multiple advantages. First, they are calculated for each individual observation, instead of representing an average on the entire sample. Second, they are applicable to all predictive methods. Third, unlike variable importance measures, variable contributions can be interpreted as negative or positive depending on their impact on the probability of an individual belonging to a given class. The typical SHAP value for an individual variable $\{i\}$ in a set of variables F is defined as follows:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left(f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)\right), \tag{1}$$

where f_S (respectively $f_{S\cup\{i\}}$) is the prediction function corresponding to the model trained using the variables contained in the subset S (respectively $S \cup \{i\}$). Similarly, x_S (respectively $x_{S\cup\{i\}}$) represents the observation of the values of variables contained in subset S (respectively $S\cup\{i\}$). Thus, the value $\phi_i \in \mathbb{R}$ is a weighted average of the marginal contributions of variable $\{i\}$ calculated when adding this variable to a model trained with a given subset of variables S.

There are different strategies for estimating SHAP values. Kernel SHAP (Lundberg and Lee, 2017) is an extension of LIME (Ribeiro et al., 2016). LIME is an additive feature attribution method that locally approximates an output $f(x)$ based on a single input x, using an explanation model g, defined as follows:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i, \tag{2}$$

where M is the number of variables, and $g(z') \approx f(h_x(z'))$ whenever $z' \approx x'$. Here, h_x is a mapping function that converts a binary vector into the original input space. Thus, x' refers to a simplified version of input x with $x = h_x(x')$. The SHAP values, denoted $\phi_i, i \in \{1, ..., M\}$, represent the parameters of the explanation model g and are jointly estimated using the Kernel SHAP method by minimising the following objective function:

$$\xi = \operatorname*{argmin}_{g\in\mathcal{G}} L(f, g, \pi_{x'}) \tag{3}$$

$$= \operatorname*{argmin}_{g\in\mathcal{G}} \sum_{z'\in Z} \left[f(h_x(z')) - g(z')\right]^2 \pi_{x'}(z'), \tag{4}$$

where $\pi_{x'}$ is a local kernel defined as follows:

$$\pi_{x'}(z') = \frac{M-1}{\binom{M}{\mid z' \mid} \mid z' \mid (M - \mid z' \mid)} \tag{5}$$

In this paper, the SHAP values are computed for the tree-based methods using the TreeSHAP method (Molnar, 2018). TreeSHAP is a more recent variant of SHAP adapted for tree-based methods. Unlike Kernel SHAP, which assumes that variables are independent, TreeSHAP uses the conditional expectation of the prediction instead. Although TreeSHAP is computationally less costly (it computes in polynomial time instead of exponential), it is known for producing unintuitive feature attributions, particularly in the case of correlated variables (Sundararajan and Najmi, 2020; Janzing et al., 2020). In the end, we decided to use this method due to its fast computation time, fully aware that independence between variables may not be guaranteed.

Once the SHAP values are estimated for each observation, we focus on respondents classified by the model as *imaginary healthy* patients. We investigate the existence of groups of individuals by performing clustering on the SHAP values of variables whose contributions, as indicated by the SHAP values, are relatively large. To identify the variables with substantial SHAP values, we adopt the following approach. First, we compute the mean SHAP value for each variable. Subsequently, the overall mean of the individual averages is calculated. Variables are retained for clustering if their average SHAP value exceeds the overall mean. We group the individuals using hierarchical clustering and use the Euclidean distance of the SHAP values as the dissimilarity measure. The number of clusters K is determined based on the silhouette score (Rousseeuw, 1987). We proceed as follows. First, we perform bootstrap re-sampling (1,000 runs) of the subset of data concerning individuals predicted $\overline{D}_{Q1-}$ by the model. For each re-sample, we perform clustering and then calculate the silhouette score for a number of K-group partitions ranging from two to eight groups. We then calculate the average of the silhouette scores over all 1,000 runs for each value of K and select the value of K that maximises the score.

3 Results

This paper focuses on the mismatch that may occur between the absence of depression and/or anxiety self-reported by individuals and their actual mental health status objectified by their MHI-5 score. This section presents the findings of the study, starting with a detailed description of the characteristics of the individuals in the observed sample. The results of the estimation using the classifiers are then presented. Finally, an interpretation of the predictions of the best classifier using the SHAP values is provided.

3.1 Study Sample and Population Characteristics

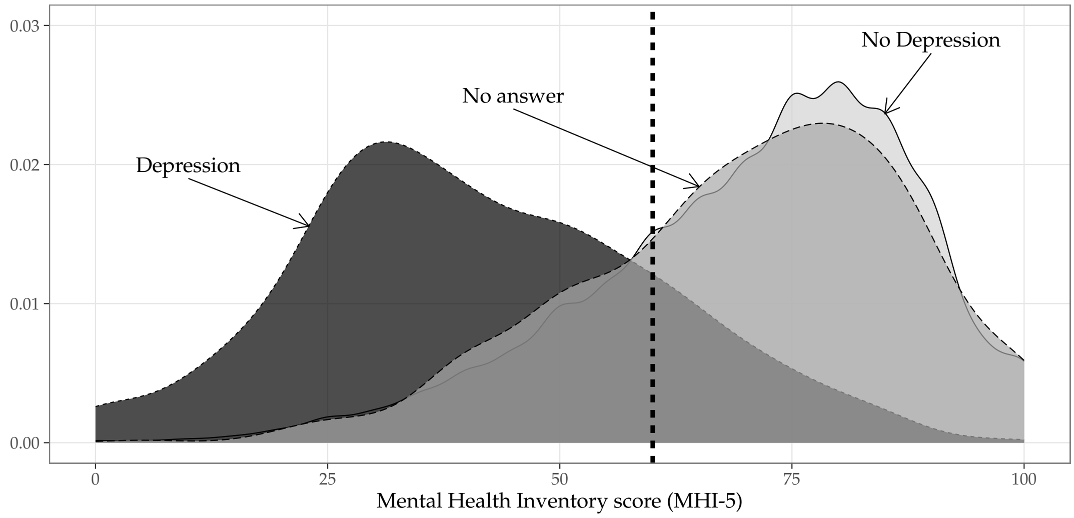
Matching the data sources presented in Section 2 results in a sample of $18,561$ individuals.[5] The whole sample is separated into three categories: individuals who reported having experienced a depressive episode in the past twelve months (705), those who reported not ($11,860$), and those for whom the information is not provided ($5,996$). The mean MHI-5 score is 41.2, 70.7, and 70.1, respectively. A broader picture of MHI-5 scores is shown in Figure 1.

Interestingly, the distribution of MHI-5 scores for individuals for whom the response to the question regarding possible depression in the past twelve months is unknown is very similar to that of individuals who reported no depressive episodes in the past year.[6] In contrast, the distribution of individuals who reported depression within the same time frame is clearly

[5] It should be noted, however, that the estimation relies on fewer observations, as explained in the current section.

[6] Comparisons between respondents who did not report mental health issues and those who did not document this point are given in Table A7 in the Appendix.

Figure 1. Density of MHI-5 Scores According to Self-reported Depression Status.

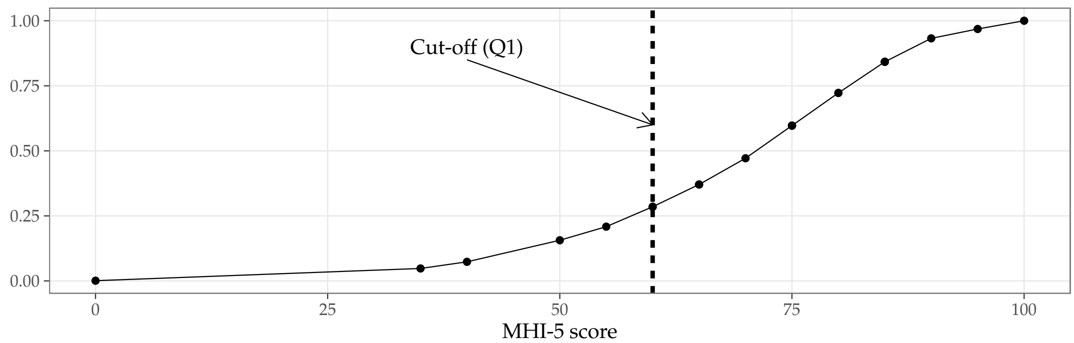


Notes: The vertical dashed line represents the first quartile of the distribution when all individuals were considered, regardless of their self-reported depression status (MHI-5 score = 60).

distinguishable from the other two. The MHI-5 scores of the respondents in this category is concentrated towards significantly lower values.

Thereafter, we focus only on respondents who reported not having experienced a depressive episode within the past twelve months. Based on their MHI-5 scores, we establish two groups: *imaginary healthy* patients, that is, those with a low MHI-5 score (denoted $\overline{D}_{Q1-}$), and others (denoted $\overline{D}_{Q1+}$). Therefore, it is necessary to select a cut-off point to establish these two groups. Figure 2 shows how the choice of this cut-off value affects the proportion of individuals designated as *imaginary healthy* patients.

Figure 2. Proportion of People with unrecognised Anxiety and/or Depressive Disorders (*Imaginary Healthy* Patients) Depending on the Value of the Cut-off Used for the MHI-5 Score.



Notes: The MHI-5 score is used to classify individuals as either healthy or not with regard to depression and/or anxiety. This classification depends on the cut-off value used for the MHI-5. This graph shows how the proportion of people with unrecognised anxiety and/or depressive disorders, that is, individuals who self-reported as not depressed while considered anxious and/or depressed according to the MHI-5 score, varies according to the cut-off. The vertical dashed line represents the first quartile of the distribution when all individuals are considered regardless of their self-reported depression status, that is, the cut-off value used in the present analysis.

Notably, there is no consensus in the literature regarding the selected values for the cut-off, as reported by Kelly et al. (2008). While some conservative values can be found (see, *e.g.*, Kroenke et al., 2005 who opt for 52, or Rist et al., 2013 and Whang et al., 2009 who use 53), much higher values can be encountered (for example Hoeymans et al., 2004 pick a value of 74). Rumpf et al. (2001) recommend using a cut-off value ranging from 60 to 70 to diagnose mood and anxiety disorders, respectively. The vertical dashed line in Figure 2 corresponds to the value we used, 60, that is, the first quartile of the distribution of MHI-5 scores over the whole sample. This corresponds to the definition used by Yamazaki et al. (2005) for moderate and severe depressive symptoms. This value also roughly corresponds to that which maximises Youden's J-statistic, as in Kelly et al. (2008). The proportion of *imaginary healthy* patients among those who report that they have not experienced depression during the previous year is 28.5% in the whole sample.

Table 1 shows the size of each sub-population defined by the intersection of self-reported depression status and MHI-5 scores.[7]

Table 1. Size of Groups of Individuals According to their MHI-5 Score and Self-reported Depression Status.

	Self-reported status					
	Depression		No Depression		Total	
MHI-5 $\leq Q_1$	572	(4.8%)	3,193	(26.9%)	3,765	(31.7%)
MHI-5 $> Q_1$	101	(0.9%)	8,006	(67.5%)	8,107	(68.2%)
Total	673	(5.6%)	11,199	(94.3%)	11,872	(100%)

Notes: This table shows the number of individuals in each subpopulation based on self-reported depression status and evaluated MHI-5 score. The MHI-5 scores range from 0 to 100, with a score of 100 representing optimal mental health. Q_1 represents the first quartile of the MHI-5 score distribution for all the respondents. The value of Q_1 is 60.

Most respondents (94.3%) reported that they had not experienced a depressive episode in the past 12 months. Among them, $3,193$ of $11,199$ (28.5%) have an MHI-5 score defined as low in this article.

The final sample contains $5,293$ individuals, which corresponds to observations with no missing values for any of the characteristics considered, as explained in Subsection 2.2.[8] Among them, $1,597$ (30.2%) have a MHI-5 score equal to or less than 60 and are labelled as *imaginary healthy* patients. Half of the participants are women, and the mean age is 48.7 ± 18.6 years. The descriptive statistics of these individuals are presented in Tables A1–A6 in the appendix.

Among the respondents who did not report any depression, women are significantly over-represented in the subsample of *imaginary healthy* patients (Table A1). There is a significant difference in the mean age between *imaginary healthy* patients (50.0 ± 18.3 years old) and respondents with MHI-5 scores higher than 60 (48.1 ± 18.7 years old). Incidentally, *imaginary healthy* patients reported poorer health status (16.0% with bad or very bad health compared to 2.4% for respondents with MHI-5 scores higher than 60). They also reported long-term conditions more often, either self-declared or documented in the Health Insurance registers (SNIIR-AM billings). The proportion of administrative and sales employees is higher in the subsample of *imaginary healthy* patients. There are significant differences in the Health Insurance scheme (*e.g.*, a higher proportion of individuals covered by the basic Universal Health Coverage among *imaginary healthy* people) and in the way respondents access Health Insurance benefits (with a greater proportion of people insured by their own among the *imaginary healthy* patients).

[7] For $6,188$ individuals among the $18,561$ aged 15 or older, the MHI-5 score cannot be computed because of missing values. In addition, 501 people did not answer questions regarding depression. Removing all of these individuals leads to $11,872$ respondents.

[8] More details on the sample selection are provided in Appendix A.

In addition, *imaginary healthy* patients live as couples less often. With regard to the binary variables indicating the self-declaration of other diseases, we observe significant differences in proportions between the two groups, the proportion of individuals who self-declare the presence of each disease (except cirrhosis) is significantly greater among the imaginary healthy patients.

Regardless of whether defined at the household or individual level, the income of *imaginary healthy* patients is significantly lower than that of respondents with MHI-5 scores higher than 60 (Table A2). There is a significant difference in household size, with smaller households among *imaginary healthy* patients.

Regarding healthcare billings from the National Health Insurance, *imaginary healthy* patients appear to be much bigger healthcare consumers than their counterparts with MHI-5 scores higher than 60 in all healthcare categories, except dentistry and optics (Table A3). Their total expenses are higher, as are those reimbursed by National Health Insurance. The gaps are much less consistent with regard to co-payments and extra fees. The variations in deductibles can be seen as indicators of differences in the healthcare services used by respondents. In the French context, all patients, except those with chronic conditions, are required to pay deductibles each time they visit a doctor, receive prescribed medications, use nursing care, or benefit from health transport. The last two rows in Table A3 reveal that, on average, *imaginary healthy* patients referred more often to general practitioners and specialists. They declared more often to renounce to healthcare provided by general practitioners or dentists because the healthcare was too distant or the appointments were too long to get (Table A4).

For respondents concerned with this issue, the big picture from the workplace is unambiguously worse for *imaginary healthy* patients (Table A5). Compared to their counterparts with MHI-5 scores higher than 60, *imaginary healthy* patients reported more often having to hurry to do their job, having little freedom, carrying heavy loads, being exposed to painful postures or to harmful or toxic products. They are also less likely to learn new things from their jobs or find help from colleagues when carrying out tasks. Finally, *imaginary healthy* people more often declared that their work required night shifts or repetitive tasks under time constraints.

Finally, social participation and interactions (participation in group activities, meetings with friends, neighbours, people within organisations, colleagues outside the workplace, or distant relatives) appear to be more restricted among *imaginary healthy* patients (Table A6). There are also significant differences in social backgrounds when looking at mothers' and fathers' levels of education, especially at both ends of the distributions.

3.2 Classification

Four types of models—RF, XGBoost, SVM, and penalised logistic regression—are estimated to classify respondents reporting no depressive episodes in the last twelve months into two categories: $\overline{D}_{Q1-}$, individuals with a low MHI-5 score (*i.e.*, *imaginary healthy* patients), and $\overline{D}_{Q1+}$, those with high MHI-5 scores. As explained in Section 2, the estimation is performed on a training sample comprising 80% of the data ($4,235$ observations and 91 explanatory variables). The values of the hyperparameters of each model obtained by repeated k-fold cross validation are reported in Table B8.

The predictive performances of the models for both training and validation samples are presented in Table 2.

The model offering the best performance in terms of sensitivity (*i.e.*, its ability to detect, among individuals defined as *imaginary healthy*, those who truly are) measured on data not used to train the model is undoubtedly XGBoost. The sensitivity, nearly twice as high as that of the other algorithms, is 69%. However, the overall capabilities of this model are inferior to the others. The accuracy on the test sample, 66.7%, is relatively lower than for the other models.

Table 2. Performances of the Classifiers.

Method	Accuracy Train	Accuracy Test	Sensitivity Train	Sensitivity Test	Specificity Train	Specificity Test	ROC-AUC Train	ROC-AUC Test
RF	0.812	0.726	0.383	0.144	0.997	0.977	0.96	0.72
XGBoost	0.676	0.667	0.666	0.690	0.680	0.658	0.74	0.71
SVM	0.780	0.721	0.348	0.226	0.967	0.935	0.86	0.72
Pen. Log. Reg.	0.740	0.749	0.276	0.301	0.941	0.942	0.75	0.73

Note: RF stands for Random Forest, XGBoost for Extreme Gradient Boosting, SVM for Support Vector Machine, and glmnet for Penalised Logistic Regression.

On the other hand, the predictive capabilities of XGBoost on the test sample are very similar to those obtained on the training sample, indicating a superior generalisation ability compared to the other models.

3.3 Interpretation with SHAP Values

In this section, we focus on the predictions made by the model offering the best performance in terms of sensitivity, XGBoost. Among the 91 explanatory variables, not all had the same relative importance in forming the score used by the model to classify observations in the target class $\overline{D}_{Q1-}$ (*imaginary healthy*). As the number of variables is high, we focus on the top 10 in terms of absolute SHAP value (Figure 3). First, the income effect appears to be the most important factor. Among the other ranked features, four are health-related: reimbursement in general practice and pharmacy, having had low back pain or neck pain over the past year. The other features are employment-related—such as having to hurry to do one's job or having very little freedom in one's job, related to social life—such as the frequency of meeting with people in organisations, or concern personal characteristics like age and gender.[9][10]

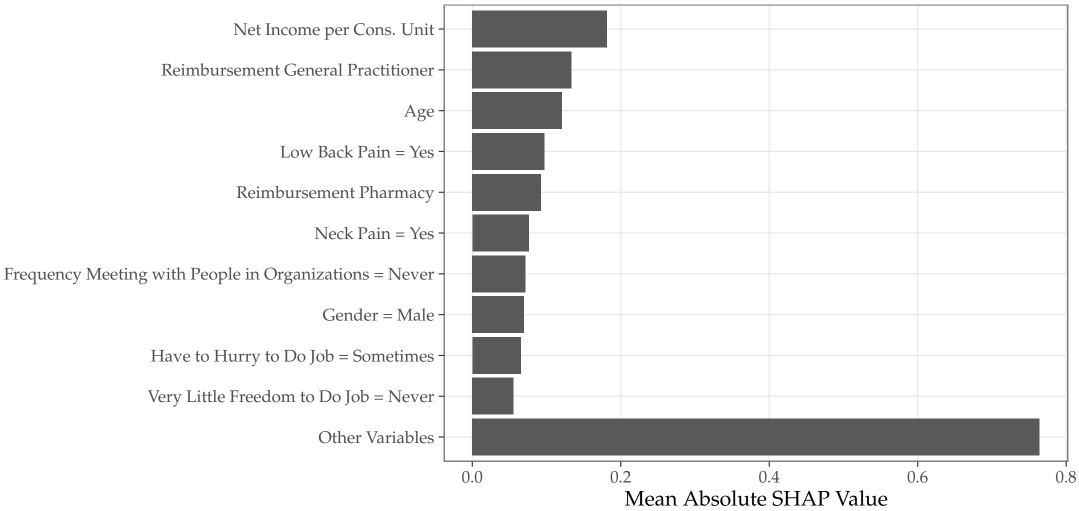
The absolute SHAP values, as shown in Figure 3, indicate nothing about the direction of the effects of the variables on the score used to classify an individual as *imaginary healthy* by the model. To do so, the signs of the SHAP values must be considered. This is proposed in Figure 4, where the effects of the main variables on the classification score for each individual according to the level of all their characteristics, are depicted. Each point represents the estimated SHAP value for each variable for each respondent. When this value is negative, the level of the variable reduces the probability of being classified as an *imaginary healthy* patient by the model. Conversely, when the SHAP value is positive, the level of the variable increases that probability. For continuous variables, an indication of the relative level of each individual variable is provided through the colour of the point representing the individual. The light grey points denote low variable values whereas black points denote high variable values. For categorical variables, the grey scale of the points varies only for ordered categorical variables. The same colour code (light gray to black) is used.

Having a net income in the lowest deciles of the distribution positively affects the probability of being classified as $\overline{D}_{Q1-}$ (Figure 4(a)). Additionally, reporting no interaction with work

[9] The variable indicating the average monthly duration of sunlight ranks 28th.

[10] Very similar results are found with the SHAP values estimated from the predictions made by the random forest. In fact, many of the top 10 variables identified with XGBoost also appear in the top 10 obtained with RF. Notably, age, gender, frequency meeting with people in organizations and freedom at work, which are in the top 10 for XGBoost, are not in the top 10 for RF. Conversely, the number of medical sessions, the deductible in pharmacy and the fact that individuals have given up a medical appointment because of the length of time it takes to obtain one, or have given up on dental care, which are in the top 10 for RF, do not appear in the top 10 for XGBoost. These detailed results can be found in the online replication ebook.

Figure 3. Variable Importance of the XGBoost Classifier According to SHAP Values.



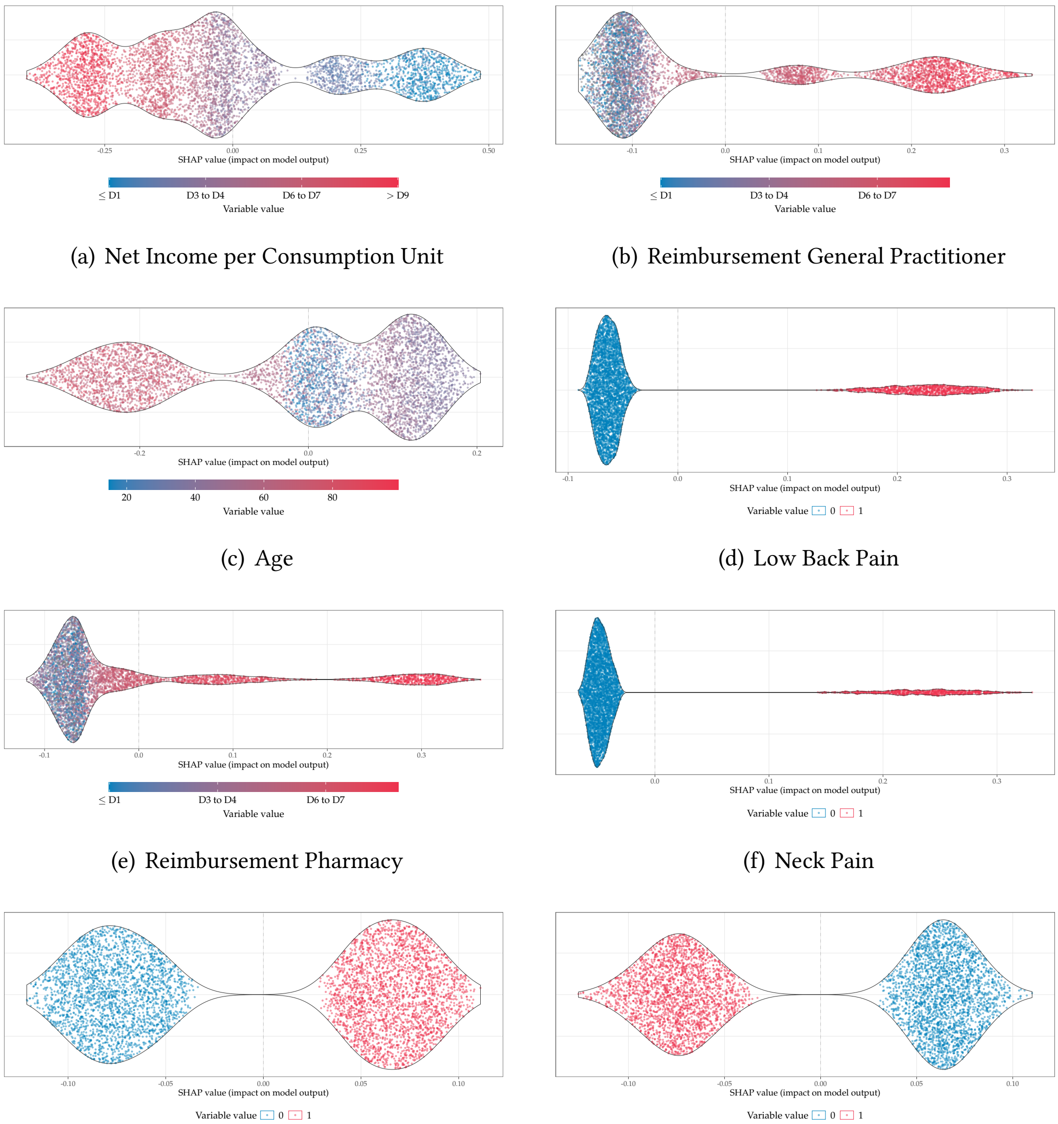
colleagues (Figure 4(g)) also positively affects the probability of being classified as $\overline{D}_{Q1-}$. These findings are consistent with the literature. A modest socio-economic status (Pena-Gralle et al., 2023), as well as a tenuous job (Asami et al., 2014), have been found to be more often associated with latent mental health problems, including depressive and anxiety disorders. Conversely, lower values for reimbursements and deductions in general practice and pharmacy (Figure 4(b), Figure 4(e)) negatively impact the probability of being classified as $\overline{D}_{Q1-}$. The healthcare consumer profile of the *imaginary healthy* identified in this article is fairly similar to that identified by other articles in the literature (Williams et al., 2017). Older age and being male (Figure 4(c) and Figure 4(h)) negatively affect the probability of being classified as $\overline{D}_{Q1-}$. Finally, having experienced lower back pain (Figure 4(d)) or neck pain (Figure 4(f)) in the past 12 months increases the likelihood that the model classifies individuals as $\overline{D}_{Q1-}$.

The SHAP values were estimated for each observation. Therefore, it is possible to decompose the contribution of each variable on an individual basis and graphically represent these contributions, as suggested by Lundberg et al. (2018).[11] We illustrate this with two examples: one where the score associated with classification by the XGBoost model as an *imaginary healthy* patient is low (first individual) and another where this score is high (second individual). We focus on the variables that most contributed, for each respective individual, to the deviation of their score from the reference score, which is the average of the scores in the sample used by TreeSHAP. The most important variable values for these individuals are reported in Table 3. The associated SHAP values are listed in order of importance in the top graph of Figure 5 for the first individual and the bottom graph for the second individual.

The score returned by the XGBoost model for the first individual is $\hat{y} = 0.217$, which is lower than the average score in the training sample ($\bar{\hat{y}} = 0.476$). The high value of net income per consumption unit (€2,400) is, according to the SHAP value, the most important factor explaining why the probability of being classified as *imaginary healthy* by the model is lower than the average prediction value. The individual's age (63 years), reimbursements of general practitioner visits (€60.4), gender (male), frequent social interactions, absence of low back pain and neck pain, most of which are below the average values in the training sample, con-

[11]We adapted the R codes of Pablo Casas and Yang Liu available on Github: https://github.com/pablo14/shap-values and https://github.com/liuyanguu/SHAPforxgboost, respectively. Our code is available in the replication ebook.

Figure 4. Impact of Variables on the Probability of Being Classified as an *Imaginary Healthy* Patient.



(a) Net Income per Consumption Unit

(b) Reimbursement General Practitioner

(c) Age

(d) Low Back Pain

(e) Reimbursement Pharmacy

(f) Neck Pain

(g) Fre. Meeting w/ People in Org. = Never

(h) Gender = Male

Note: The plots are ordered by variable importance with respect to average absolute SHAP values. Each dot represents a single individual. For points with a negative abscissa, the variable of interest had a downward effect on the probability of being classified as an *imaginary healthy* patient ($\overline{D}_{Q1-}$). For quantitative variables, the colour of the points depends on the level of the variable of interest, ranging from light grey (low values) to black (high values). For the net income per consumption unit variable, the grey scale ranges according to the empirical quantiles of the variable.

tribute to lowering the probability of the model predicting the individual as *imaginary healthy*. Conversely, the amounts for pharmacy reimbursement (€396), which is higher than the sample average, prevent the probability from decreasing further. For the second individual, the high amounts for pharmacy reimbursement (€1,813.33) and general practitioner reimbursement (€206.20), along with the low income per consumption unit (€900), are the three main factors explaining why the score returned by the model, $\hat{y} = 0.714$, is higher than the average score. The advanced age (79 years) and the absence of low back pain reduces the probability of the model classifying the second individual as *imaginary healthy*. Unlike the first individual, the gender variable contributes to increasing the score. For the second individual, the fact of never meeting other people in organisations, contrary to what might be expected, decreases the probability of being classified as an *imaginary healthy* patient.

Table 3. Characteristics for the Most Important Variables for Two Individuals According to their SHAP Values.

Variable	First Individual ($\hat{y} = 0.217$)		Second Individual ($\hat{y} = 0.714$)		Train Set Average Value
	Value	In top 10	Value	In top 10	
Net Income per Cons. Unit	2,400	✓	900	✓	1,610.26
Age	63	✓	79	✓	48.56
Reimbursement General Practitioner	60.4	✓	206.2	✓	87.52
Gender = Male	1	✓	0	✓	0.48
Freq. Meet. w/ People in Org.: At least once a week	1	✓	0		0.17
Freq. Meet. w/ People in Org.: Never	0	✓	0	✓	0.51
Low Back Pain = Yes	0	✓	0	✓	0.2
Participation in Group Activities: No	0	✓	1		0.63
Neck Pain: Yes	0	✓	0		0.15
Reimbursement Pharmacy	396.26	✓	1,818.33	✓	364.21
Long-term condition (Self-declared): No	1		0	✓	0.81
No. Medical Sessions General Pract.	4		10	✓	4.64
Waiver Appointment Delay Too Long: No	1		0	✓	0.68

Note: The predicted score by the XGBoost model is $\hat{y} = 0.217$ for the first individual and $\hat{y} = 0.747$ for the second. The average score in the reference sample is $\bar{\hat{y}} = 0.476$. The column "Value" indicates the value of the individual's characteristics, while the column "In top 10" shows whether each variable is among the top 10 most important variables in explaining the deviation of the individual's score from the average score of the reference sample. The last column gives the average value of the variables in the reference set used by TreeSHAP, *i.e.*, the train set.

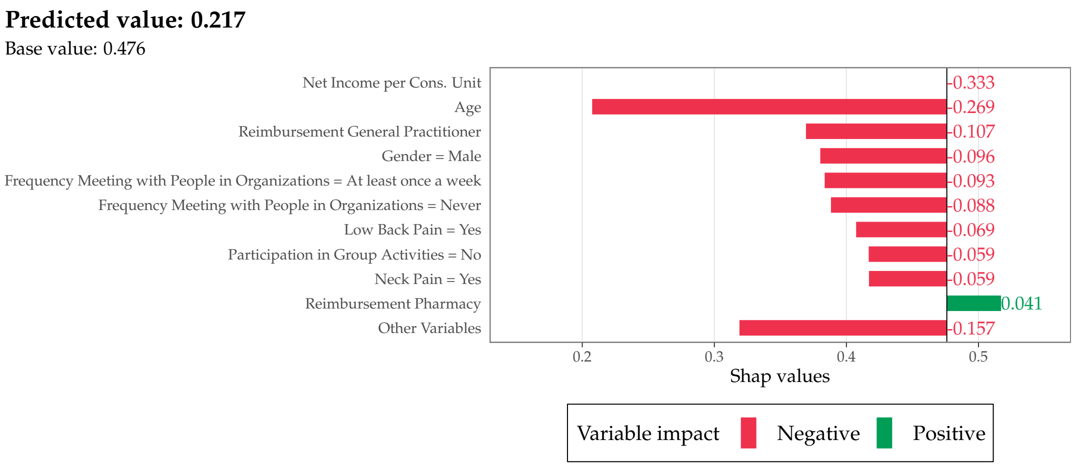
Rather than looking at this type of results for each observation in the dataset, it is possible, as explained in Lundberg et al. (2018), to stack the contributions of each individual vertically and place the individual contributions on a horizontal axis to produce what the authors call a 'force plot.' We rely on this type of representation and focus exclusively on individuals predicted as *imaginary healthy* ($\overline{D}_{Q1-}$) by the XGBoost classifier. Only the contributions of the most important variables for each individual are reported,[12] the effects of the other variables being summed and labelled as "Other." The observations are ordered based on similarities in their SHAP values. These similarities were obtained from ascending hierarchical classification. According to the silhouette score (Section 2), three profiles emerge among individuals classified $\overline{D}_{Q1-}$. Figure 6 shows the details of these profiles in terms of the average SHAP values of each variable observed in each cluster.[13] Table 4 complements the figure and shows the means and standard deviations of the top 10 variables for each cluster as well as for the set of individuals predicted as *imaginary healthy* and for the entire sample.

The first cluster, comprising 626 individuals with a 48.9% true positive rate, is characterized by elderly individuals (mean age 66 years) with substantial healthcare consumption. The probability of being classified as imaginary healthy in this group is primarily driven by lower income levels (€$1,194 \pm 687$ compared to the sample average of €$1,610 \pm 1008$) and advanced age. However, a notable counterbalancing factor emerges: pharmaceutical consumption. Substantial pharmaceutical expenditures (€$1,795 \pm 3322$ versus €$361 \pm 1,432$ in the overall sample) significantly decrease the likelihood of being classified as imaginary healthy. The second and

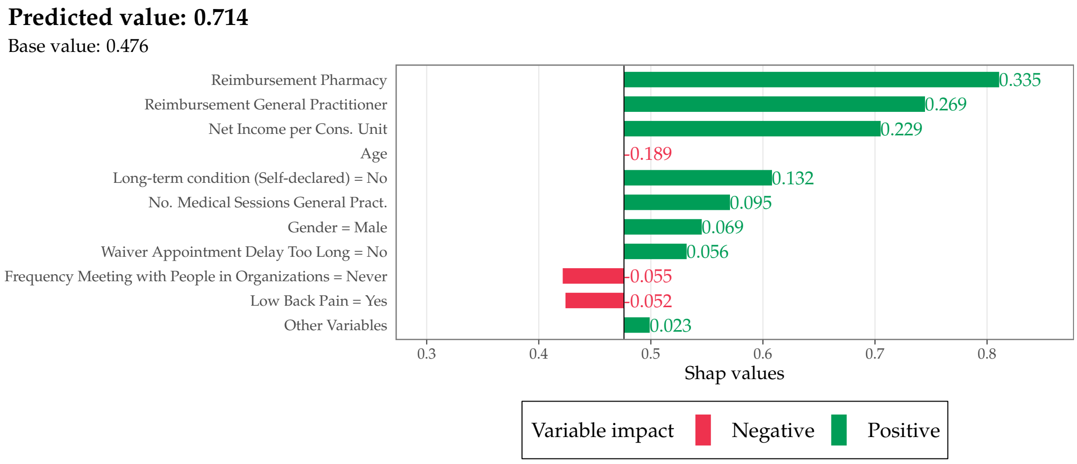
[12] These are the eleven variables with the highest average absolute SHAP values for this subset of individuals predicted as $\overline{D}_{Q1-}$. The number of variables defining the top is chosen such that it corresponds to a group of variables with similar average absolute SHAP values. Beyond the eleventh variable, the average absolute SHAP values become less significant.

[13] Refer to the online ebook, Chapter 8, for individual-level representations.

Figure 5. Decomposition of the Contribution of the Most Influential Variables to the Prediction Deviation of Being Classified as an *Imaginary Healthy* Patient from the Baseline Value.



(a) Individual with a low predicted score

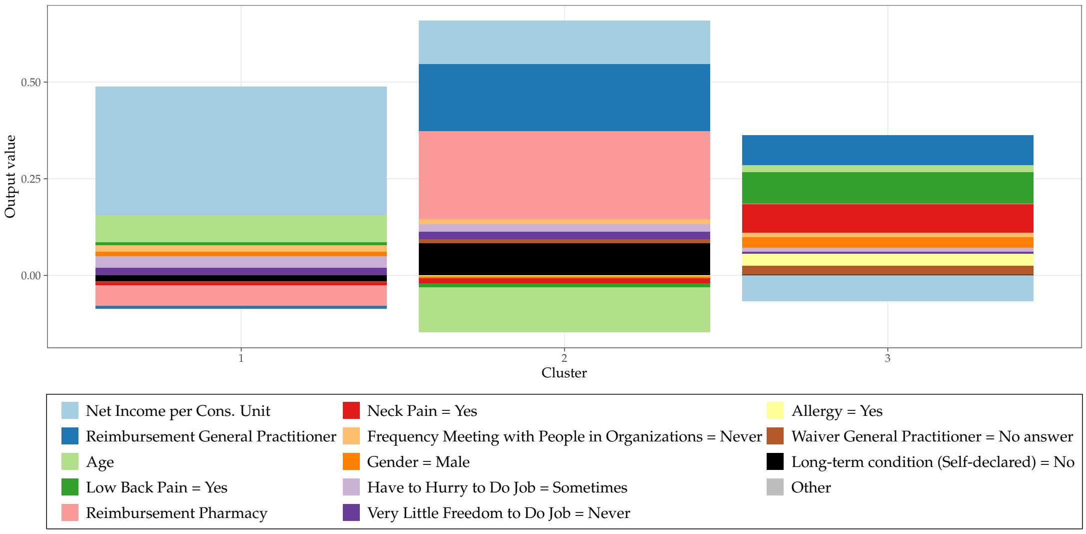


(b) Individual with a high predicted score

Note: The baseline value corresponds to the average probability of being classified by the preferred model as an *imaginary healthy* patient, ($\overline{D}_{Q1-}$) in the dataset.

largest cluster ($n = 1,077$) presents a more complex interplay of predictive factors, with a 48% true positive rate. Several characteristics positively influence the classification as imaginary healthy: the pharmaceutical and general practitioner reimbursements (€131 $\pm$ 120 compared to €89 $\pm$ 113 in the sample average), higher income levels (€$1,641 \pm 691$), and the general absence of long-term conditions (79% report none). Conversely, the age profile (mean 49 ± 16 years) emerges as a negative predictor of imaginary health classification. The third cluster ($n = 566$), showing a lower true positive rate of 43.8%, reveals particularly interesting determinants of imaginary health classification. Interestingly, the absence of reported neck pain or low back pain increases the likelihood of being classified as imaginary healthy, with 75.6% reporting no low back pain and 86.9% reporting no neck pain. The moderate general practitioner reimbursement levels (€86 ± 109) also positively influence this classification. However,

Figure 6. Decomposition of the Effect of the Most Important Variables on the Probability of Being Predicted as an *Imaginary Healthy* Patient for Individuals Predicted as Such.



Note: The reported contribution is relative to the baseline value (the average probability of being classified by the preferred model as an *imaginary healthy* patient ($\overline{D}_{Q1-}$) in the dataset) centered at zero.

the low level of income (€643 $\pm$ 192) substantially decreases the probability of being classified as imaginary healthy. Lastly, this cluster is characterized by high social isolation (66.8% report no organizational participation).

Table 4. Average Person in Each Cluster and in the Samples.

	Cluster 1	Cluster 2	Cluster 3	Pred. Imaginary	Pred. Healthy	Entire Sample
	$n = 626$	$n = 1,077$	$n = 566$	$n = 2,269$	$n = 3,024$	$n = 5,293$
Imaginary healthy	306(48.9%)	517(48.0%)	248(43.8%)	1071(47.2%)	526(17.4%)	1597(30.2%)
Accuracy	48.9%	48.0%	43.8%	47.2%	82.6%	67.4%
Net Income per Cons. Unit	1193.67 (687.48)	1640.95 (691.14)	643.45 (192.42)	1268.72 (728.7)	1865.21 (1108.24)	1609.51 (1008.13)
Reimbursement GP	228.85 (169.81)	131.15 (119.72)	85.79 (108.71)	146.79 (143.51)	46.41 (48.95)	89.44 (112.53)
Age	66.15 (15.23)	49.29 (16.48)	38.39 (12.82)	51.22 (18.39)	46.8 (18.49)	48.7 (18.58)
Low Back Pain	No (81.6%)	No (50.3%)	No (75.6%)	No (65.3%)	No (90.4%)	No (79.6%)
Reimbursement Pharmacy	1794.9 (3321.96)	306.26 (669.55)	99.92 (178.35)	665.5 (1937.75)	133.41 (805.81)	361.5 (1431.61)
Neck Pain	No (88%)	No (57.9%)	No (86.9%)	No (73.5%)	No (94.3%)	No (85.4%)
Freq. Meet. w/ People in Org.	Never (64.2%)	Never (59.4%)	Never (66.8%)	Never (62.6%)	Never (42.9%)	Never (51.3%)
Gender	Female (51%)	Female (71.7%)	Female (62.9%)	Female (63.8%)	Male (56.4%)	Female (52.2%)
Have to Hurry to Do Job	No answer (85.8%)	No answer (37.5%)	No answer (71.6%)	No answer (59.3%)	No answer (46.2%)	No answer (51.8%)
Very Little Freedom to Do Job	No answer (85.9%)	No answer (37.8%)	No answer (72.1%)	No answer (59.6%)	No answer (46.4%)	No answer (52.1%)
Allergy	No (86.7%)	No (69.5%)	No (83.9%)	No (77.8%)	No (91.1%)	No (85.4%)
Waiver GP	No (81%)	No (85.1%)	No (64%)	No (78.7%)	No (74%)	No (76%)
Long-term condition (Self-declared)	Yes (74.1%)	No (79.4%)	No (89.9%)	No (67.2%)	No (90.5%)	No (80.5%)

Notes: This table shows the representative person in each of the three clusters in the set of individuals predicted as *imaginary healthy*, in the set of individuals predicted as healthy, and in the entire sample. For numerical variables, the within-cluster means are reported (with standard errors between brackets), whereas for categorical variables, the within-cluster mode is provided (the proportions are given in brackets). The first rows of the table give the number of *imaginary healthy* and the corresponding proportion in each sample, and the accuracy of the XBG model in that sample. GP stands for general practitioners.

4 Discussion

In this article, we portrayed the *imaginary healthy* patients using tree-based methods. *Imaginary healthy* patients were defined as those who did not report spontaneously having or having had recent mental health problems but whose MHI-5 scores were very poor. These methods have been preferred over more classical parametric estimations because of the tricky issue of multiple potential endogeneity between our outcome variable (non-diagnosed or non-reported mental health trouble) and several associated variables that can be explanatory to, as well as explained by, the outcome variable (occupation, marital status, long-term condition, income, occupational position, working conditions, healthcare expenditures, social participation, and interactions) in a cross-sectional design that makes it difficult to identify any causal relationship. In this regard, our results provide evidence of significant characteristics associated with *imaginary healthy* patients, taking into account the potential interactions between characteristics and nonlinearity.

A big picture of *imaginary healthy* patients identifies respondents whose health status is poorer than that of their counterparts who appears to correctly assess their mental health status. This poorer health status seems to be supported by self-assessed health status and chronic conditions, objectivised by healthcare consumption billings and chronic conditions traceable by the National Health Insurance. Regarding the global health conditions of the respondents, the misperception by *imaginary healthy* patients of their true mental health status appears to be a latent personal trait associated with actual and consequent pathologies, just as a mental/psychological comorbidity. In this figure, the working conditions are at the forefront. *Imaginary healthy* patients exhibited much less favourable working conditions than respondents who declared no mental health issues and whose MHI-5 scores were higher than 60. Because of the questions from the job content questionnaire questionnaire (Karasek, 1985; Karasek et al., 1998), it appears that *imaginary healthy* patients have much more demanding occupational activities, exposing them to sustained work rates and painful working conditions, potentially including the handling of heavy loads and harmful substances, with less decision-making flexibility and fewer opportunities to get help from co-workers to complete tasks (Williams et al., 2017; Pena-Gralle et al., 2023; Rugulies et al., 2023).

In addition to their relatively professional isolation (Kouvonen et al., 2008), *imaginary healthy* people appear to be socially isolated, and they tend to have less social participation and interactions. At this stage, our article does not allow us to determine the causal direction of the relationship between social and/or professional deprivation and anxiety-depressive disorders (Rugulies et al., 2023; Johnson, 1991). Living less often as a couple, *imaginary healthy* patients may be less able to find resources to rely on within the couple. Taken together, these characteristics are also associated with an overall less comfortable economic condition for *imaginary healthy* patients than for their healthy counterparts.

Identifying what fuels the lack of insight that this study sheds light on into *imaginary healthy* people is not self-evident. It remains unclear whether the unconsciousness of their actual psychological disorders, called anosognosia in Alzheimer's and brain-injured patients (Starkstein et al., 2006; Orfei et al., 2009), is a disease *per se* for the selected respondents to the ESPS survey, or the expression of a denial, which is an expected psychological defence mechanism (Goldbeck, 1997). How patients assess their actual health status, including their mental health status and the potential healthcare needs, may vary dramatically (Prins et al., 2008). However, the reliability of self-assessment of health is questionable. The assessments may not be the same over time for the same actual health status because of the different lists of questions used to document the self-reported health status (Crossley and Kennedy, 2002). Self-assessed health may also involve heterogeneity bias, which has been widely documented in the literature. Robust empirical evidence has been found regarding systematic differences in response styles across countries (Jürges, 2007) or social classes (Jürges, 2008), or depending on the position in the income distribution (Etilé and Milcent, 2006), gender, or age (Lindeboom and van Doorslaer, 2004). This heterogeneity bias may be due to intrinsic discrepancies in how individuals view different health statuses differently, depending on their social class, socioeconomic status, and culture.

Self-reported health can also be manipulated by respondents when their economic situations rely heavily on it. Bound (1991) and Kerkhofs and Lindeboom (1995), among others, demonstrated how self-assessment of health can be biased by considering social benefits. Additionally, we cannot reject the fact that these results may be due to differential item functioning (Teresi, 2006; Teresi and Fleishman, 2007; Knott et al., 2017). Facing the same health situation, respondents may not use the same items as those proposed in the questionnaire, possibly because of social, economic, or demographic factors (*e.g.*, gender, education, ethnicity, deprivation). For all these reasons, given that our contribution does not specifically address the issue of potential response bias or differential item functioning, although it could be considered as a means of detecting them, the results obtained in this paper should be greeted with caution.

In our article, we use the MHI-5 score for the detection of mental health conditions (anxiety and depressive symptoms) to check for self-report of any recent illness (over the past twelve months), including depression.

Of course, we cannot totally crowd out that respondents did not declare any depressive episode, backed by neither official medico-administrative recognition of it nor specific healthcare consumption that we would have identified in our dataset, whereas the MHI-5 score captured emerging mental health issues. Unfortunately, this grey area cannot be illuminated using the data available to us. In this respect, it would certainly have been useful to know how respondents' consumption of healthcare had changed in the twelve months following the survey. In addition, some *imaginary healthy* patients may turn out to be not as imaginary as we claim, simply because their MHI-5 score would report anxiety in the last four weeks when they have rightly declared no depression in the last twelve months. Nonetheless, anxiety is known to be most often comorbid with depression (Kalin, 2020), both being associated with chronic medical illnesses (Katon et al., 2007), what we globally find with our data, especially in Cluster 1. Finally, the psychometric properties of the MHI-5 score can also be questioned, particularly its ability to identify depressive disorders. Thereon, the MHI-5 is presented as '*a good screener for mood disorder in the general population, with high sensitivity and specificity*' and '*when patients are screened for major depression and/or dysthymia, the three-item MHI-d[epression] is as good as the full MHI-5*' (Cuijpers et al., 2009). In that respect, we checked the robustness of our findings in two distinct directions. First, when restricted to the MHI-3 (only the three dimensions for depression in the MHI-5 score, Yamazaki et al., 2005), the results remain globally the same. The predictive performance of the model is slightly better overall, but mainly because of its better ability to predict not being *imaginary healthy*. Second, when the self-assessed health is substituted with the question about a depressive episode in the last twelve months and crossed with the MHI-5 score to identify *imaginary healthy* patients, *imaginary healthy* patients are still associated with poor social and family interactions and working conditions (low-skilled, low-support, and unempowering jobs), as in Borrell et al. (2004).[14]

One of the potential limitations of our study lies in the selection process for the final sample on which the estimates were based (Figure A.1). Starting with an initial dataset of $19,940$ individuals, we excluded individuals under the age of 15 years old (n = 1,379) to focus on an adult population. This exclusion inherently limits the generalization of our results to older individuals and not to younger ones. Additionally, individuals with missing values on the MHI-5 score ($n = 6,188$) were excluded, which may introduce bias if those with missing MHI-5 scores differ from the others, possibly along unobserved psychological or socioeconomic dimensions. Furthermore, we excluded individuals who had affirmative or missing values on the variable that self-reports presence of depression ($n = 1,174$), as both the latter and the

[14] In the online ebook, **Part IV Robustness checks: MHI-3** and **Part V Robustness checks: SAH** respectively give the codes and propose replications of all the results detailed in the body of the paper when substituting the MHI-5 score with the MHI-3 score and when substituting the reported depression with the self-assessed health respectively, regarding descriptive statistics (chapters 9 and 10, and chapters 16 and 17 respectively) , classification results (chapters 11 and 12, and chapters 17 and 18 respectively) and SHAP values (chapters 13 and 14, and chapters 19 and 20 respectively).

MHI-5 score are crucial for constructing our variable of interest (being an imaginary healthy patient or not). This exclusion may contribute to selection bias if individuals with depressive symptoms or missing responses differ in important ways from the rest of the sample. Lastly, individuals with missing values on several predictors were excluded ($n = 5,906$) to maintain data completeness in the rest of the analysis, a step which may limit the representativeness of the sample if those with missing predictors have distinct characteristics. All of these selection steps result in a final sample of $5,305$ individuals and may affect the generalization of our results, even if for some of those steps, even though for some of these stages we systematically checked the potential specificities of the individuals excluded from the sample against those still included on the basis of common observable variables. Future research could consider alternative approaches, such as multiple imputation for missing data, to minimize potential bias and enhance representativeness.

Nonetheless, with the help of clusters, our results help to identify three distinct populations whose profiles deserve the attention of anyone contributing to the defence of population health, whatever the level (centralised or decentralised) and the motivation (public health or economic interest). The individuals in the first cluster are characterised by a dramatically high consumption of healthcare (consultations with GPs and reimbursed drugs are about 5 times and 3 times higher than the mean amount in the entire sample, respectively), in association with an outstanding prevalence of chronic issues. In this cluster, the prediction for respondents to be *imaginary healthy* is positively determined by their modest individual disposable income, their advanced age, their male gender and their hardly emancipating working conditions. The under-diagnosed and/or under-reported mental disorders of these people mainly appear as masked co-morbidities or complications of otherwise recognised somatic health problems (Verhaak et al., 2005; Scott et al., 2016). In many respects (age, individual disposable income, absence of chronic issue, consultations with GPs), respondents in the second cluster have the average profile of the whole sample. Yet, compared to respondents in the first cluster, they are more likely to be female and, while mainly reporting no chronic issue, they significantly evoke much more often back and neck pain, without incurring higher pharmacy costs. Low back pain mainly contributes to predict them as *imaginary healthy*, and so do consultations with GPs, individual disposable income, the absence of recognised chronic issue or not-so-adverse working conditions, among others. Because of their possible association with mood and anxiety disorders (Von Korff and Simon, 1996; Bair et al., 2003), the actual source of the pain complaints of these individuals must be questioned. The profile of respondents in the third cluster is more difficult to relate to any recognised or masked health issue. Indeed, at first sight, their young age, their low use of healthcare (whether in terms of visits to the GP or use of medication) and the fact that they rarely report chronic health problems suggest that they are objectively healthy people. But they are also economically deprived, with an individual disposable income less than half the sample mean. Even if scarcer than for the two other clusters, pain complaints are the main predictors of their classification as *imaginary healthy*, as well as consultations with GPs and personal disposable income. Regarding respondents in the third cluster, individual but also neighbourhood deprivation can constitute a strong signal for potential mental health concerns (Patel et al., 2007; Visser et al., 2021).

It is tricky to make recommendations to the public health authorities based on our results. However, it seems to us that, whatever their level, the stakeholders concerned by the predictive characteristics of the risk that the individuals in each cluster are *imaginary healthy* people, have levers that can help them raise awareness of *imaginary healthy* people of the health problems they are likely to develop. With regard to the first cluster, the high level of healthcare expenditure in relation to diagnosed chronic illnesses should alert both compulsory health insurers and supplementary insurers to the particular risk of developing associated anxiety-depressive disorders. The use of algorithms (Choong et al., 2024; Hernandez et al., 2021) to identify health issues, including anxiety and depression (Priya et al., 2020; Zulfiker et al., 2021), is now widespread and provides a tractable tool for the identification of insurees to whom proposing suitable screening questionnaires on the basis of which it would possible, if they reveal unrecognised mental disorders, to provide imaginary health patients with appropriate preventive action. Beyond the basic public health interest that those actions may target, those interventions can reveal to be cost-efficient (Gräfe et al., 2020) for the compulsory national as well as the supplementary health insurers (Wanni Arachchige Dona et al., 2021). The exercise is certainly much more difficult to carry out with individuals belonging to the second cluster, because of the absence of prominent health problems to refer to, but low back and neck pain. If these complaints are addressed to a healthcare professional, there is a good chance that they will make a fair assessment of the risk of developing anxiety-depressive disorders. If these complaints are not verbalized, it is to be hoped that they can be spotted in one circumstance or another. The workplace can be one of them, for example during surveys on quality of life at work and working conditions (Bender and Farvolden, 2008). Once again, as long as employers wish their employees healthy for altruistic as well as productive concerns, the identification of *imaginary health* employees can prove to be a cost-efficient action, not so expensive to carry out and likely to prevent the costs associated with absenteism and presenteism. Finally, the third group is certainly the least specific to be dealt with in a particular prevention programme. However, it does have two key characteristics that need to be addressed: individuals belonging to that cluster are mainly young and described by very modest economic conditions. The prevention of anxiety and depressive disorders must be particularly focused on young adults, as part of national schemes that can make use of the health framework of the school and university systems, which in the case of France should be given a new lease of life (Nauphal et al., 2023). Furthermore, social programmes that target poverty may have the unexpected but welcome side-effect of reducing or preventing the risk of being *imaginary healthy* which individuals of modest socio-economic status would not be aware of and which would escape epidemiological surveillance (Ridley et al., 2020). In other contexts, programmes aimed at reducing poverty using cash transfers proved to be efficient regarding the population mental health (McGuire et al., 2022).

5 Conclusion

Based on matching the 2012 French National Representative Health and Social Protection Survey and annual healthcare consumption billings from the French health insurance record, this article examines factors related to unrecognised anxiety and depressive troubles. The mental status related to anxiety and depressive disorders are assessed using the MHI-5 score. Crossing this score with the self-reported perception of depression of the people in the survey allows us to define two categories of people among those reporting no depressive episodes in the last twelve months: those who were unaware that they had anxiety or depressive disorders, which we call *imaginary healthy*, and others in whom these disorders did not occur. Once these two types of people are labelled, we train machine learning algorithms, to predict whether a person belongs to the *imaginary healthy* category based on their personal characteristics, healthcare consumption, and environmental, social, and family conditions. While the preferred classifier, XGBoost, exhibits modest predictive capabilities, it identifies factors strongly associated with the inability to recognise the presence of anxiety and depressive disorders. The predictions from our machine learning algorithm corroborate the empirical evidence provided by descriptive statistics by highlighting the existence of profiles related to the onset of anxiety and depressive disorders. More specifically, a gender effect consistent with epidemiological knowledge regarding the prevalence of anxiety and depressive disorders in the general population is observed alongside an income effect associated with people with unrecognised mental health problems and low personal disposable income. We also observe a massive effect expressing the overall arduousness of the occupation (feeling of not having enough time, low decision latitude in the job), and occupation type. The effect of reduced social participation also stands out. More marginally and probably in connection with previous factors, specific medical consumption (pharmacy and consultations with general practitioners) is found to be associated with the inability to recognise the presence of anxiety or depressive disorders. Of course, as advocated by Rugulies et al. (2023), the causal transmission channels between undiagnosed mental health issues and social and labour outcomes require more comprehensive investigation, which our article is not able to do on the basis of cross-sectional data, and what could be achieved with the help of the challenging techniques of causal random forests initiated by Wager and Athey (2018) on longitudinal datasets.

Be that as it may, for the public health decision-maker, the compulsory and/or supplementary health insurer, the MHI-5 score combined with any self-assessed health indicator represents an inexpensive, easy-to-use and reliable instrument for screening for unrecognised or undiagnosed mental disorders in the general population, which can be administered by a health professional in a practice, in the workplace or at home using a self-administered questionnaire. Predictions based on machine learning methods, whether they identify true or false positives, can help to identify patients whose latent health needs need to be met quickly (Haghish and Czajkowski, 2024).

References

Aragonès, E., Piñol, J. L. and Labad, A. (2006). The overdiagnosis of depression in non-depressed patients in primary care. *Family Practice* 23: 363–368, doi:10.1093/fampra/cmi120. 2

Asami, Y., Goren, A. and Okumura, Y. (2014). Work Productivity Loss with Depression, Diagnosed and Undiagnosed, among Employed Respondents in an Internet-Based Survey Conducted in Japan. *Value in Health* 17: A463, doi:10.1016/j.jval.2014.08.1289. 2, 12

Bair, M. J., Robinson, R. L., Katon, W. and Kroenke, K. (2003). Depression and Pain Comorbidity: A Literature Review. *Archives of Internal Medicine* 163: 2433–2445, doi:10.1001/archinte.163.20.2433. 19

Beck, A., Crain, L. A., Solberg, L. I., Unützer, J., Maciosek, M. V., Whitebird, R. R. and Rossom, R. C. (2014). The effect of depression treatment on work productivity. *The American Journal of Managed Care* 20: e294–301. 2

Ben-Hur, A., Horn, D., Siegelmann, H. T. and Vapnik, V. (2002). Support vector clustering. *Journal of Machine Learning Research* 2: 125–137. 37

Bender, A. and Farvolden, P. (2008). Depression and the workplace: A progress report. *Current Psychiatry Reports* 10: 73–79, doi:10.1007/s11920-008-0013-6. 20

Berndt, E. R., Finkelstein, S. N., Greenberg, P. E., Howland, R. H., Keith, A., Rush, A. J., Russell, J. and Keller, M. B. (1998). Workplace performance effects from chronic depression and its treatment. *Journal of Health Economics* 17: 511–535, doi:10.1016/S0167-6296(97)00043-X. 2

Borrell, C., Muntaner, C., Benach, J. and Artazcoz, L. (2004). Social class and self-reported health status among men and women: what is the role of work organisation, household material standards and household labour? *Social Science & Medicine (1982)* 58: 1869–1887, doi:10.1016/S0277-9536(03)00408-8. 18

Bound, J. (1991). Self-Reported Versus Objective Measures of Health in Retirement Models. *The Journal of Human Resources* 26: 106–138, doi:10.2307/145718. 17

Breiman, L. (2001). Random Forests. *Machine Learning* 45: 5–32, doi:10.1023/A:1010933404324. 6, 37

Broadhead, W. E., Blazer, D. G., George, L. K. and Tse, C. K. (1990). Depression, disability days, and days lost from work in a prospective epidemiologic survey. *JAMA* 264: 2524–2528, doi:10.1001/jama.1990.03450190056028. 2

Cacheda, F., Fernandez, D., Novoa, F. J. and Carneiro, V. (2019). Early Detection of Depression: Social Network Analysis and Random Forest Techniques. *Journal of Medical Internet Research* 21: e12554, doi:10.2196/12554. 37

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. New York, NY, USA: Association for Computing Machinery, 785–794, doi:10.1145/2939672.2939785. 37

Choong, C., Brnabic, A., Chinthammit, C., Ravuri, M., Terrell, K. and Kan, H. (2024). Applying machine learning approaches for predicting obesity risk using us health administrative claims database. *BMJ Open Diabetes Research and Care* 12, doi:10.1136/bmjdrc-2024-004193. 20

Crossley, T. F. and Kennedy, S. (2002). The reliability of self-assessed health status. *Journal of Health Economics* 21: 643–658, doi:10.1016/S0167-6296(02)00007-3. 17

Crown, W. H. (2015). Potential application of machine learning in health outcomes research and some statistical cautions. *Value in Health* 18: 137 – 140, doi:10.1016/j.jval.2014.12.005. 37

Cuijpers, P., Smits, N., Donker, T., ten Have, M. and de Graaf, R. (2009). Screening for mood and anxiety disorders with the five-item, the three-item, and the two-item mental health inventory. *Psychiatry Research* 168: 250–255, doi:https://doi.org/10.1016/j.psychres.2008.05.012. 18

Doupe, P., Faghmous, J. and Basu, S. (2019). Machine learning for health services researchers. *Value in Health* 22: 808 – 815, doi:10.1016/j.jval.2019.02.012. 2

Druss, B. G., Wang, P. S., Sampson, N. A., Olfson, M., Pincus, H. A., Wells, K. B. and Kessler, R. C. (2007). Understanding Mental Health Treatment in Persons Without Mental Diagnoses: Results From the National Comorbidity Survey Replication. *Archives of General Psychiatry* 64: 1196–1203, doi:10.1001/archpsyc.64.10.1196. 2

Egede, L. E. (2007). Failure to Recognize Depression in Primary Care: Issues and Challenges. *Journal of General Internal Medicine* 22: 701–703, doi:10.1007/s11606-007-0170-z. 2

Etilé, F. and Milcent, C. (2006). Income-related reporting heterogeneity in self-assessed health: evidence from France. *Health Economics* 15: 965–981, doi:10.1002/hec.1164. 17

Eveleigh, R., Grutters, J., Muskens, E., Oude Voshaar, R., Weel, C. van, Speckens, A. and Lucassen, P. (2014). Cost-utility analysis of a treatment advice to discontinue inappropriate long-term antidepressant use in primary care. *Family Practice* 31: 578–584, doi:10.1093/fampra/cmu043. 2

Falagas, M. E., Vardakas, K. Z. and Vergidis, P. I. (2007). Under-diagnosis of common chronic diseases: prevalence and impact on human health. *International Journal of Clinical Practice* 61: 1569–1579, doi:10.1111/j.1742-1241.2007.01423.x. 2

Freeling, P., Rao, B. M., Paykel, E. S., Sireling, L. I. and Burton, R. H. (1985). Unrecognised depression in general practice. *Br Med J (Clin Res Ed)* 290: 1880–1883, doi:10.1136/bmj.290.6485.1880. 2

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33, doi:10.18637/jss.v033.i01. 38

Goldbeck, R. (1997). Denial in physical illness. *Journal of Psychosomatic Research* 43: 575–593, doi:10.1016/S0022-3999(97)00168-2. 17

Gräfe, V., Moritz, S. and Greiner, W. (2020). Health economic evaluation of an internet intervention for depression (deprexis), a randomized controlled trial. *Health Economics Review* 10: 19, doi:10.1186/s13561-020-00273-0. 20

Haghish, E. F. and Czajkowski, N. (2024). Reconsidering False Positives in Machine Learning Binary Classification Models of Suicidal Behavior. *Current Psychology* 43: 10117–10121, doi:10.1007/s12144-023-05174-z. 21

Hart, S. (1989). Shapley Value. In Eatwell, J., Milgate, M. and Newman, P. (eds), *Game Theory*, The New Palgrave. London: Palgrave Macmillan UK, 210–216, doi:10.1007/978-1-349-20181-5_25. 6

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc. 2

Hernandez, E. J. M., Tingzon, I., Ampil, L. and Tiu, J. (2021). Identifying chronic disease patients using predictive algorithms in pharmacy administrative claims: an application in rheumatoid arthritis. *Journal of Medical Economics* 24: 1272–1279, doi:10.1080/13696998.2021.1999132, pMID: 34704871. 20

Herr, M., Sirven, N., Grondin, H., Pichetti, S. and Sermet, C. (2017). Frailty, polypharmacy, and potentially inappropriate medications in old people: findings in a representative sample of the French population. *European Journal of Clinical Pharmacology* 73: 1165–1172, doi:10.1007/s00228-017-2276-5. 3

Heyden, J. H. A. Van der, Gisle, L., Hesse, E., Demarest, S., Drieskens, S. and Tafforeau, J. (2009). Gender differences in the use of anxiolytics and antidepressants: a population based study. *Pharmacoepidemiology and Drug Safety* 18: 1101–1110, doi:10.1002/pds.1827. 2

Higgins, E. S. (1994). A review of unrecognized mental illness in primary care. Prevalence, natural history, and efforts to change the course. *Archives of Family Medicine* 3: 908–917, doi:10.1001/archfami.3.10.908. 2

Hilton, M. F., Scuffham, P. A., Vecchio, N. and Whiteford, H. A. (2010). Using the interaction of mental health symptoms and treatment status to estimate lost employee productivity. *Australian and New Zealand Journal of Psychiatry* 44: 151–161, doi:10.3109/00048670903393605. 2

Hoeymans, N., Garssen, A. A., Westert, G. P. and Verhaak, P. F. (2004). Measuring mental health of the Dutch population: a comparison of the GHQ-12 and the MHI-5. *Health and Quality of Life Outcomes* 2: 23, doi:10.1186/1477-7525-2-23. 4, 9

Janzing, D., Minorics, L. and Blöbaum, P. (2020). Feature relevance quantification in explainable AI: A causal problem. In *International Conference on artificial intelligence and statistics*. PMLR, 2907–2916. 7

Johnson, T. P. (1991). Mental health, social relations, and social selection: A longitudinal analysis. *Journal of Health and Social Behavior* 32: 408–423. 17

Jureidini, J. and Tonkin, A. (2006). Overuse of Antidepressant Drugs for the Treatment of Depression. *CNS Drugs* 20: 623–632, doi:10.2165/00023210-200620080-00002. 2

Jürges, H. (2007). True health vs response styles: exploring cross-country differences in self-reported health. *Health Economics* 16: 163–178, doi:10.1002/hec.1134. 17

Jürges, H. (2008). Self-assessed health, reference levels and mortality. *Applied Economics* 40: 569–582, doi:10.1080/00036840500447823. 17

Kalin, N. H. (2020). The critical relationship between anxiety and depression. *American Journal of Psychiatry* 177: 365–367, doi:10.1176/appi.ajp.2020.20030305, pMID: 32354270. 18

Karasek, R. (1985). *Job Content Questionnaire and user's guide*. University of Massachusetts Lowell, Department of Work Environment: Lowell. 4, 17

Karasek, R., Brisson, C., Kawakami, N., Houtman, I., Bongers, P. and Amick, B. (1998). The job content questionnaire (JCQ): An instrument for internationally comparative assessments of psychosocial job characteristics. *Journal of Occupational Health Psychology* 3: 322–355, doi:10.1037/1076-8998.3.4.322. 4, 17

Katon, W., Lin, E. H. and Kroenke, K. (2007). The association of depression and anxiety with medical symptom burden in patients with chronic medical illness. *General Hospital Psychiatry* 29: 147–155, doi:https://doi.org/10.1016/j.genhosppsych.2006.11.005. 18

Kelly, M. J., Dunstan, F. D., Lloyd, K. and Fone, D. L. (2008). Evaluating cutpoints for the MHI-5 and MCS using the GHQ-12: a comparison of five different methods. *BMC Psychiatry* 8: 10, doi:10.1186/1471-244X-8-10. 9

Kent, S. T., McClure, L. A., Crosson, W. L., Arnett, D. K., Wadley, V. G. and Sathiakumar, N. (2009). Effect of sunlight exposure on cognitive function among depressed and non-depressed participants: a regards cross-sectional study. *Environmental Health* 8, doi:10.1186/1476-069x-8-34. 4

Kerkhofs, M. and Lindeboom, M. (1995). Subjective health measures and state dependent reporting errors. *Health Economics* 4: 221–235, doi:10.1002/hec.4730040307. 17

Kessler, R. C., Barber, C., Birnbaum, H. G., Frank, R. G., Greenberg, P. E., Rose, R. M., Simon, G. E. and Wang, P. (1999). Depression In The Workplace: Effects On Short-Term Disability. *Health Affairs* 18: 163–171, doi:10.1377/hlthaff.18.5.163. 2

Kim, S. Y., Bang, M., Wee, J. H., Min, C., Yoo, D. M., Han, S.-M., Kim, S. and Choi, H. G. (2021). Short- and long-term exposure to air pollution and lack of sunlight are associated with an increased risk of depression: A nested case-control study using meteorological data and national sample cohort data. *Science of The Total Environment* 757: 143960, doi:10.1016/j.scitotenv.2020.143960. 4

Klinkman, M. S., Coyne, J. C., Gallo, S. and Schwenk, T. L. (1998). False Positives, False Negatives, and the Validity of the Diagnosis of Major Depression in Primary Care. *Archives of Family Medicine* 7: 451, doi:10.1001/archfami.7.5.451. 2

Knott, R. J., Lorgelly, P. K., Black, N. and Hollingsworth, B. (2017). Differential item functioning in quality of life measurement: An analysis using anchoring vignettes. *Social Science & Medicine* 190: 247–255, doi:10.1016/j.socscimed.2017.08.033. 18

Koopmans, P. C., Roelen, C. a. M. and Groothoff, J. W. (2008). Sickness absence due to depressive symptoms. *International Archives of Occupational and Environmental Health* 81: 711–719, doi:10.1007/s00420-007-0243-7. 2

Kouvonen, A., Oksanen, T., Vahtera, J., Stafford, M., Wilkinson, R., Schneider, J., Väänänen, A., Virtanen, M., Cox, S. J., Pentti, J. et al. (2008). Low workplace social capital as a predictor of depression: the finnish public sector study. *American Journal of Epidemiology* 167: 1143–1151, doi:10.1093/aje/kwn067. 17

Kroenke, C. H., Bennett, G. G., Fuchs, C., Giovannucci, E., Kawachi, I., Schernhammer, E., Holmes, M. D. and Kubzansky, L. D. (2005). Depressive Symptoms and Prospective Incidence of Colorectal Cancer in Women. *American Journal of Epidemiology* 162: 839–848, doi:10.1093/aje/kwi302. 9

Lamu, A. N. and Olsen, J. A. (2016). The relative importance of health, income and social relations for subjective well-being: An integrative analysis. *Social Science & Medicine* 152: 176 – 185, doi:https://doi.org/10.1016/j.socscimed.2016.01.046. 2

Lim, D., Sanderson, K. and Andrews, G. (2000). Lost productivity among full-time workers with mental disorders. *The Journal of Mental Health Policy and Economics* 3: 139–146, doi:10.1002/mhp.93. 2

Lindeboom, M. and Doorslaer, E. van (2004). Cut-point shift and index shift in self-reported health. *Journal of Health Economics* 23: 1083–1099, doi:10.1016/j.jhealeco.2004.01.002. 17

Lundberg, S. M. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. and Garnett, R. (eds), *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 4765–

4774. 6

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K.-W., Newman, S.-F., Kim, J. and Lee, S.-I. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering* 2: 749–760, doi:10.1038/s41551-018-0304-0. 12, 14

Luppa, M., Heinrich, S., Angermeyer, M. C., König, H.-H. and Riedel-Heller, S. G. (2008). Health-care costs associated with recognized and unrecognized depression in old age. *International Psychogeriatrics* 20: 1219–1229, doi:10.1017/S1041610208007680. 2

McGuire, J., Kaiser, C. and Bach-Mortensen, A. M. (2022). A systematic review and meta-analysis of the impact of cash transfers on subjective well-being and mental health in low- and middle-income countries. *Nature Human Behaviour* 6: 359–370, doi:10.1038/s41562-021-01252-z. 20

McQuaid, J. R., Stein, M. B., Laffaye, C. and McCahill, M. E. (1999). Depression in a Primary Care Clinic: the Prevalence and Impact of an Unrecognized Disorder. *Journal of Affective Disorders* 55: 1–10, doi:10.1016/S0165-0327(98)00191-8. 2

Melchior, M., Ferrie, J. E., Alexanderson, K., Goldberg, M., Kivimaki, M., Singh-Manoux, A., Vahtera, J., Westerlund, H., Zins, M. and Head, J. (2009). Using sickness absence records to predict future depression in a working population: prospective findings from the GAZEL cohort. *American Journal of Public Health* 99: 1417–1422, doi:10.2105/AJPH.2008.142273. 2

Mojtabai, R. and Olfson, M. (2011). Proportion Of Antidepressants Prescribed Without A Psychiatric Diagnosis Is Growing. *Health Affairs* 30: 1434–1442, doi:10.1377/hlthaff.2010.1024. 2

Molnar, C. (2018). *Interpretable Machine Learning (Second Edition)*. Leanpub. 7

Nauphal, M., Ward-Ciesielski, E. and Eustis, E. H. (2023). Preventing anxiety and depression in emerging adults: A case for targeting help-seeking intentions and behaviors. *Journal of Prevention and Health Promotion* 4: 112–143, doi:10.1177/26320770221124802. 20

Orfei, M. D., Caltagirone, C. and Spalletta, G. (2009). The Evaluation of Anosognosia in Stroke Patients. *Cerebrovascular Diseases* 27: 280–289, doi:10.1159/000199466. 17

Patel, V., Flisher, A. J., Hetrick, S. and McGorry, P. (2007). Mental health of young people: a global public-health challenge. *The Lancet* 369: 1302–1313, doi:10.1016/S0140-6736(07)60368-7, publisher: Elsevier. 19

Pearson, R., Pisner, D., Meyer, B., Shumake, J. and Beevers, C. G. (2019). A machine learning ensemble to predict treatment outcomes following an Internet intervention for depression. *Psychological Medicine* 49: 2330–2341, doi:10.1017/S003329171800315X. 37

Pena-Gralle, A. P. B., Talbot, D., Trudel, X., Milot, A., Gilbert-Ouimet, M., Lavigne-Robichaud, M., Ndjaboué, R., Lesage, A., Lauzier, S., Vézina, M. et al. (2023). Socioeconomic inequalities, psychosocial stressors at work and physician-diagnosed depression: Time-to-event mediation analysis in the presence of time-varying confounders. *Plos One* 18: e0293388, doi:10.1371/journal.pone.0293388. 12, 17

Pierre, A. and Jusot, F. (2017). The likely effects of employer-mandated complementary health insurance on health coverage in France. *Health Policy* 121: 321–328, doi:10.1016/j.healthpol.2016.12.004. 3

Prins, M. A., Verhaak, P. F. M., Bensing, J. M. and Meer, K. van der (2008). Health beliefs and perceived need for mental health care of anxiety and depression—The patients' perspective explored. *Clinical Psychology Review* 28, doi:10.1016/j.cpr.2008.02.009. 17

Priya, A., Garg, S. and Tigga, N. P. (2020). Predicting anxiety, depression and stress in modern life using machine learning algorithms. *Procedia Computer Science* 167: 1258–1267, doi:https://doi.org/10.1016/j.procs.2020.03.442, international Conference on Computational Intelligence and Data Science. 20

Ribeiro, M. T., Singh, S. and Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv:1602.04938 [cs, stat]* . 6

Ridley, M., Rao, G., Schilbach, F. and Patel, V. (2020). Poverty, depression, and anxiety: Causal evidence and mechanisms. *Science* 370: eaay0214, doi:10.1126/science.aay0214. 20

Rist, P. M., Schürks, M., Buring, J. E. and Kurth, T. (2013). Migraine, headache, and the risk of depression: Prospective cohort study. *Cephalalgia* 33: 1017–1025, doi:10.1177/0333102413483930. 9

Rizzo, J. A., Abbott, T. A. and Pashko, S. (1996). Labour productivity effects of prescribed medicines for chronically ill workers. *Health Economics* 5: 249–265, doi:10.1002/(SICI)1099-1050(199605)5:3<249::AID-HEC203>3.0.CO;2-A. 2

Rost, K., Zhang, M., Fortney, J., Smith, J., Coyne, J. and Smith, G. R. (1998). Persistently poor outcomes of undetected major depression in primary care. *General Hospital Psychiatry* 20: 12–20, doi:10.1016/s0163-8343(97)00095-9. 2

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20: 53–65, doi:10.1016/0377-0427(87)90125-7. 7

Rugulies, R., Aust, B., Greiner, B. A., Arensman, E., Kawakami, N., LaMontagne, A. D. and Madsen, I. E. (2023). Work-related causes of mental health conditions and interventions for their improvement in workplaces. *The Lancet* 402: 1368–1381, doi:10.1016/S0140-6736(23)00869-3. 17, 21

Rumpf, H.-J., Meyer, C., Hapke, U. and John, U. (2001). Screening for mental health: validity of the MHI-5 using DSM-IV Axis I psychiatric disorders as gold standard. *Psychiatry Research* 105: 243–253, doi:10.1016/s0165-1781(01)00329-8. 4, 9

Scott, K. M., Lim, C., Al-Hamzawi, A., Alonso, J., Bruffaerts, R., Almeida, J. M. Caldas-de, Florescu, S., Girolamo, G. de, Hu, C., Jonge, P. de, Kawakami, N., Medina-Mora, M. E., Moskalewicz, J., Navarro-Mateu, F., O'Neill, S., Piazza, M., Posada-Villa, J., Torres, Y. and Kessler, R. C. (2016). Association of Mental Disorders With Subsequent Chronic Physical Conditions: World Mental Health Surveys From 17 Countries. *JAMA Psychiatry* 73: 150, doi:10.1001/jamapsychiatry.2015.2688. 19

Sheehan, D. V. (2004). Depression: underdiagnosed, undertreated, underappreciated. *Managed Care (Langhorne, Pa.)* 13: 6–8. 2

Simon, G. E., Barber, C., Birnbaum, H. G., Frank, R. G., Greenberg, P. E., Rose, R. M., Wang, P. S. and Kessler, R. C. (2001). Depression and Work Productivity: The Comparative Costs of Treatment Versus Nontreatment. *Journal of Occupational and Environmental Medicine* 43: 2, doi:10.1097/00043764-200101000-00002. 2

Starkstein, S. E., Jorge, R., Mizrahi, R. and Robinson, R. G. (2006). A diagnostic formulation for anosognosia in Alzheimer's disease. *Journal of Neurology, Neurosurgery & Psychiatry* 77: 719–725, doi:10.1136/jnnp.2005.085373. 17

Stewart, W. F., Ricci, J. A., Chee, E., Hahn, S. R. and Morganstein, D. (2003). Cost of Lost Productive Work Time Among US Workers With Depression. *JAMA* 289: 3135–3144, doi:10.1001/jama.289.23.3135. 2

Sundararajan, M. and Najmi, A. (2020). The many Shapley values for model explanation. In *International conference on machine learning*. PMLR, 9269–9278. 7

Teresi, J. A. (2006). Different approaches to differential item functioning in health applications. Advantages, disadvantages and some neglected topics. *Medical Care* 44: S152–170, doi:10.1097/01.mlr.0000245142.74628.ab. 18

Teresi, J. A. and Fleishman, J. A. (2007). Differential Item Functioning and Health Assessment. *Quality of Life Research* 16: 33–42, doi:10.1007/s11136-007-9184-6. 18

Thorsen, S. V., Rugulies, R., Hjarsbech, P. U. and Bjorner, J. B. (2013). The predictive value of mental health for long-term sickness absence: the Major Depression Inventory (MDI) and the Mental Health Inventory (MHI-5) compared. *BMC Medical Research Methodology* 13: 115, doi:10.1186/1471-2288-13-115. 4

Verhaak, P. F., Heijmans, M. J., Peters, L. and Rijken, M. (2005). Chronic disease and mental disorder. *Social Science & Medicine* 60: 789–797, doi:https://doi.org/10.1016/j.socscimed.2004.06.012. 19

Visser, K., Bolt, G., Finkenauer, C., Jonker, M., Weinberg, D. and Stevens, G. W. (2021). Neighbourhood deprivation effects on young people's mental health and well-being: A systematic review of the literature. *Social Science & Medicine* 270: 113542, doi:https://doi.org/10.1016/j.socscimed.2020.113542. 19

Von Korff, M. and Simon, G. (1996). The relationship between pain and depression. *British Journal of Psychiatry* 168: 101–108, doi:10.1192/S0007125000298474. 19

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* 113: 1228–1242, doi:10.1080/01621459.2017.1319839. 21

Wang, R., Liu, Y., Xue, D., Yao, Y., Liu, P. and Helbich, M. (2019). Cross-sectional associations between long-term exposure to particulate matter and depression in china: The mediating effects of sunlight, physical activity, and neighborly reciprocity. *Journal of Affective Disorders* 249: 8–14, doi:10.1016/j.jad.2019.02.007. 4

Wanni Arachchige Dona, S., Angeles, M. R., Hall, N., Watts, J. J., Peeters, A. and Hensher, M. (2021). Impacts of chronic disease prevention programs implemented by private health insurers: a systematic review. *BMC Health Services Research* 21: 1222, doi:10.1186/s12913-021-07212-7. 20

Ware, J. E., Kosinski, M. and Dewey, J. E. (2001). *How to score version 2 of the SF-36 health survey: (standard & acute forms) ; [SF-36v2]*. Lincoln, RI: QualityMetric, 3rd ed. 4

Whang, W., Kubzansky, L. D., Kawachi, I., Rexrode, K. M., Kroenke, C. H., Glynn, R. J., Garan, H. and Albert, C. M. (2009). Depression and Risk of Sudden Cardiac Death and Coronary Heart Disease in Women. *Journal of the American College of Cardiology* 53: 950–958, doi:10.1016/j.jacc.2008.10.060. 9

Williams, S. Z., Chung, G. S. and Muennig, P. A. (2017). Undiagnosed depression: A community diagnosis. *SSM - Population Health* 3: 633–638, doi:https://doi.org/10.1016/j.ssmph.2017.07.012. 12, 17

Yamazaki, S., Fukuhara, S. and Green, J. (2005). Usefulness of five-item and three-item Mental Health Inventories to screen for depressive symptoms in the general population of Japan. *Health and Quality of Life Outcomes* 3: 48, doi:10.1186/1477-7525-3-48. 9, 18

Zulfiker, M. S., Kabir, N., Biswas, A. A., Nazneen, T. and Uddin, M. S. (2021). An in-depth analysis of machine learning approaches to predict depression. *Current Research in Behavioral Sciences* 2: 100044, doi:https://doi.org/10.1016/j.crbeha.2021.100044. 20

Appendix

A Sample Composition

Matching the different parts of the ESPS survey results in a dataset of $19,940$ individuals. However, there are many missing values in it. In addition, the set of individuals does not correspond to the people whom we intend to target. The flowchart in Figure A1 shows the successive removals of observations leading to the final sample used for the estimation of the model.

First, we exclude $1,379$ people under 15. Then, we remove $6,188$ individuals for whom it is not possible to compute the MHI-5 score (these people answered two or less of the five questions used to compute the MHI-5 score). As the aim of the article is to study *imaginary healthy* people, we only include people who declared that they had not experienced a depressive episode in the last 12 months. This leads to the separation of 501 people who did not answer the question on depression, and 673 people who reported having had a depressive episode in the previous year.

We then remove a few people due to missing values for the following variables: coverage by the compulsory national health insurance (48 observations), health insurance scheme (4 observations) and occupation (10 observations). For the latter variable, it should be noted that individuals who have never worked are included in the sample.

Information on household disposable income is missing for $1,561$ people, which we *de facto* remove from the final sample. Unfortunately, data on health care consumption (of any kind among the many categories provided) are not available for $4,270$ people, which we also remove. We then remove one respondent for whom the questions on economic and social living conditions are missing.

Some parts of the questionnaire were administered only to a subsample of respondents. When an individual's responses were absent because they were not prompted to answer the questions, rather than by their choice to withhold answers, we categorise the responses as 'No Answer.' This decision was made to prevent the exclusion of the sample. Finally, we discard 12 individuals for whom couple status is not available, even after trying to infer their status. In the raw sample, the couple status is not available for 292 people. We make the following assumptions for these individuals : for all individuals who were 15 years old (this question was not asked to people younger than 15 years), we assume that they do not live in a couple. For people older than 15 years and living alone, we assume they are not living as a couple either. For the remaining 12 individuals, we do not infer anything about this matter and discard them from the analysis.

Descriptive statistics for the study sample are presented in Tables A1 to A6. Each table shows the means and standard deviations for numerical variables, and the proportions of each level for qualitative variables. The first column provides statistics for people who self-reported that they had not experienced a depressive episode in the last twelve months. The second and third columns provide the same statistics for those with low MHI-5 scores (below the first quartile of the distribution) and those with high MHI-5 scores, respectively. The fourth column provides a test p-value depending on the type of variables (analysis of variance for continuous variables, χ^2-test for nominal variables, and Kruskal-Wallis rank sum test for ordinal variables). Bold values indicate significant differences (p-value < 0.05).

Table A1. Descriptive Statistics of Individual Characteristics.

Variable	**Self-reported No depression**	**MHI-5 $\leq Q1$**	**MHI-5 $> Q1$**	p-value
Expenses				
MHI-5 Score	69.97 (±17.93)	47.7 (±11.96)	79.59 (±9.58)	$\mathbf{< 10^{-3}}$
Self-Assessed Health Condition				
Not reported	22 (0.42%)	5 (0.31%)	17 (0.46%)	
Very Bad or Bad	344 (6.5%)	256 (16.03%)	88 (2.38%)	
Very Good, Good or Fairly Good	4927 (93.09%)	1336 (83.66%)	3591 (97.16%)	$\mathbf{< 10^{-3}}$
Age	48.7 (±18.58)	50.04 (±18.28)	48.11 (±18.67)	$\mathbf{< 10^{-3}}$
Gender				
Female	2765 (52.24%)	950 (59.49%)	1815 (49.11%)	
Male	2528 (47.76%)	647 (40.51%)	1881 (50.89%)	$\mathbf{< 10^{-3}}$
Couple				
No	1830 (34.57%)	636 (39.82%)	1194 (32.31%)	
Yes	3463 (65.43%)	961 (60.18%)	2502 (67.69%)	$\mathbf{< 10^{-3}}$

Continued on next page

Table A1 – Continued from previous page

Variable	Self-reported No depression	MHI-5 $\leq Q1$	MHI-5 $> Q1$	p-value
Marital Status				
Married, Civil Union	2861 (54.65%)	780 (48.96%)	2081 (57.14%)	
Divorced, separated	437 (8.35%)	197 (12.37%)	240 (6.59%)	
Widowed	336 (6.42%)	129 (8.1%)	207 (5.68%)	
Lives in a marriage or concubinage	600 (11.46%)	181 (11.36%)	419 (11.5%)	
Single	996 (19.03%)	306 (19.21%)	690 (18.95%)	
Does not know	1 (0.02%)	0 (0.00%)	1 (0.03%)	
Refuses to answer	4 (0.08%)	0 (0.00%)	4 (0.11%)	$< 10^{-3}$
Professional Status				
Public employee	854 (16.13%)	280 (17.53%)	574 (15.53%)	
Private employee	3153 (59.57%)	974 (60.99%)	2179 (58.96%)	
Other	737 (13.92%)	207 (12.96%)	530 (14.34%)	
No answer	549 (10.37%)	136 (8.52%)	413 (11.17%)	**0.005**
Social Security				
Yes (own)	4554 (86.04%)	1402 (87.79%)	3152 (85.28%)	
Yes (third party)	739 (13.96%)	195 (12.21%)	544 (14.72%)	**0.018**
Social Security System				
The general scheme (Cnamts)	3959 (74.8%)	1211 (75.83%)	2748 (74.35%)	
Public service	294 (5.55%)	80 (5.01%)	214 (5.79%)	
The local Alsace-Moselle scheme	187 (3.53%)	67 (4.2%)	120 (3.25%)	
The basic Universal Health Coverage	151 (2.85%)	76 (4.76%)	75 (2.03%)	
The agricultural scheme	340 (6.42%)	83 (5.2%)	257 (6.95%)	
The self-employed scheme	338 (6.39%)	76 (4.76%)	262 (7.09%)	
Other (Student, abroad, other)	19 (0.36%)	2 (0.13%)	17 (0.46%)	
Does not know	5 (0.09%)	2 (0.13%)	3 (0.08%)	$< 10^{-3}$
Occupation				
Farmer	160 (3.02%)	38 (2.38%)	122 (3.3%)	
Craftsman, trader	292 (5.52%)	80 (5.01%)	212 (5.74%)	
Executive and intellectual profession	745 (14.08%)	184 (11.52%)	561 (15.18%)	
Intermediate occupation	931 (17.59%)	255 (15.97%)	676 (18.29%)	
Administrative employee	713 (13.47%)	265 (16.59%)	448 (12.12%)	
Commercial employee	650 (12.28%)	236 (14.78%)	414 (11.2%)	
Skilled worker	824 (15.57%)	248 (15.53%)	576 (15.58%)	
Unskilled worker	418 (7.9%)	149 (9.33%)	269 (7.28%)	
Inactive having never worked	560 (10.58%)	142 (8.89%)	418 (11.31%)	$< 10^{-3}$
Long-term condition (Self-declared)				
Yes	1018 (19.23%)	430 (26.93%)	588 (15.91%)	
No	4262 (80.52%)	1161 (72.7%)	3101 (83.9%)	
Does not know	13 (0.25%)	6 (0.38%)	7 (0.19%)	$< 10^{-3}$
Long-term condition (SNIIRAM)				
Yes	1019 (19.25%)	408 (25.55%)	611 (16.53%)	
No	4274 (80.75%)	1189 (74.45%)	3085 (83.47%)	$< 10^{-3}$
Asthma (n = 5,293)				
Yes	359 (6.78%)	165 (10.33%)	194 (5.25%)	$< 10^{-3}$
No	4934 (93.22%)	1432 (89.67%)	3502 (94.75%)	
Bronchitis (n = 5,293)				
Yes	306 (5.78%)	159 (9.96%)	147 (3.98%)	$< 10^{-3}$
No	4987 (94.22%)	1438 (90.04%)	3549 (96.02%)	
Heart Attack (n = 5,293)				
Yes	34 (0.64%)	19 (1.19%)	15 (0.41%)	**0.002**
No	5259 (99.36%)	1578 (98.81%)	3681 (99.59%)	
Artery Disease (n = 5,293)				
Yes	103 (1.95%)	49 (3.07%)	54 (1.46%)	$< 10^{-3}$
No	5190 (98.05%)	1548 (96.93%)	3642 (98.54%)	
Hypertension (n = 5,293)				
Yes	631 (11.92%)	254 (15.9%)	377 (10.2%)	$< 10^{-3}$
No	4662 (88.08%)	1343 (84.1%)	3319 (89.8%)	
Stroke (n = 5,293)				
Yes	28 (0.53%)	14 (0.88%)	14 (0.38%)	**0.037**
No	5265 (99.47%)	1583 (99.12%)	3682 (99.62%)	
Osteoarthritis (n = 5,293)				
Yes	707 (13.36%)	298 (18.66%)	409 (11.07%)	$< 10^{-3}$
No	4586 (86.64%)	1299 (81.34%)	3287 (88.93%)	
Low Back Pain (n = 5,293)				
Yes	1078 (20.37%)	458 (28.68%)	620 (16.77%)	$< 10^{-3}$
No	4215 (79.63%)	1139 (71.32%)	3076 (83.23%)	

Continued on next page

Table A1 – Continued from previous page

Variable	Self-reported No depression	MHI-5 $\leq Q1$	MHI-5 $> Q1$	p-value
Neck Pain (n = 5,293)				
Yes	774 (14.62%)	353 (22.1%)	421 (11.39%)	$\mathbf{< 10^{-3}}$
No	4519 (85.38%)	1244 (77.9%)	3275 (88.61%)	
Diabetes (n = 5,293)				
Yes	445 (8.41%)	223 (13.96%)	222 (6.01%)	$\mathbf{< 10^{-3}}$
No	4848 (91.59%)	1374 (86.04%)	3474 (93.99%)	
Allergy (n = 5,293)				
Yes	773 (14.6%)	313 (19.6%)	460 (12.45%)	$\mathbf{< 10^{-3}}$
No	4520 (85.4%)	1284 (80.4%)	3236 (87.55%)	
Cirrhosis (n = 5,293)				
Yes	7 (0.13%)	4 (0.25%)	3 (0.08%)	0.253
No	5286 (99.87%)	1593 (99.75%)	3693 (99.92%)	
Urinary Incontinence (n = 5,293)				
Yes	222 (4.19%)	117 (7.33%)	105 (2.84%)	$\mathbf{< 10^{-3}}$
No	5071 (95.81%)	1480 (92.67%)	3591 (97.16%)	

Notes: The MHI-5 scores range from 0 to 100, with a score of 100 representing optimal mental health. $N = 5,293$. The first column gives the characteristics of the sample of people who self-reported not having depression, regardless of the MHI-5 score. The second and third columns consider only those people whose MHI-5 score was less than or equal to $Q1 = 60$ and strictly greater than $Q1 = 60$, respectively. The different characteristics are expressed in terms of the mean $\pm$ standard deviation for continuous variables, and in terms of absolute and relative frequencies for nominal and ordinal variables. The fourth column provides a test p-value, depending on the type of variable (analysis of variance for continuous variables, χ^2-test for nominal variables, and Kruskal-Wallis rank sum test for ordinal variables). Bold values indicate significant differences (p-value < 0.05).

Table A2. Descriptive Statistics of Household and Regional Characteristics.

Variable	Self-reported No depression	MHI-5 $\leq Q1$	MHI-5 $> Q1$	p-value
Expenses				
Household size	2.84 ($\pm$1.44)	2.73 ($\pm$1.45)	2.89 ($\pm$1.43)	$\mathbf{< 10^{-3}}$
Income (*note used in the model*)	2855.27 ($\pm$1893.64)	2432.83 ($\pm$1495.36)	3039.56 ($\pm$2015.88)	$\mathbf{< 10^{-3}}$
Net Income per Cons. Unit	1609.51 ($\pm$1008.13)	1424.22 ($\pm$805.62)	1689.57 ($\pm$1074.24)	$\mathbf{< 10^{-3}}$
Regional Characteristics				
Zoning in Urban Areas				
Major urban cluster (and its crown)	3680 (69.53%)	1046 (65.5%)	2634 (71.27%)	
Medium/small urban cluster (and its crown)	697 (13.17%)	205 (12.84%)	492 (13.31%)	
Spaces outside the area of influence of cities	238 (4.5%)	61 (3.82%)	177 (4.79%)	
No answer	678 (12.81%)	285 (17.85%)	393 (10.63%)	$\mathbf{< 10^{-3}}$
Region				
Région parisienne	739 (13.96%)	221 (13.84%)	518 (14.02%)	
Champagne-Ardenne	148 (2.8%)	49 (3.07%)	99 (2.68%)	
Picardie	200 (3.78%)	64 (4.01%)	136 (3.68%)	
Haute-Normandie	145 (2.74%)	30 (1.88%)	115 (3.11%)	
Centre	194 (3.67%)	61 (3.82%)	133 (3.6%)	
Basse-Normandie	141 (2.66%)	45 (2.82%)	96 (2.6%)	
Bourgogne	148 (2.8%)	46 (2.88%)	102 (2.76%)	
Nord-Pas-de-Calais	376 (7.1%)	134 (8.39%)	242 (6.55%)	
Lorraine	245 (4.63%)	82 (5.13%)	163 (4.41%)	
Alsace	154 (2.91%)	52 (3.26%)	102 (2.76%)	
Franche-Comté	115 (2.17%)	28 (1.75%)	87 (2.35%)	
Pays de la Loire	344 (6.5%)	98 (6.14%)	246 (6.66%)	
Bretagne	281 (5.31%)	66 (4.13%)	215 (5.82%)	
Poitou-Charentes	178 (3.36%)	44 (2.76%)	134 (3.63%)	
Aquitaine	278 (5.25%)	79 (4.95%)	199 (5.38%)	
Midi-Pyrénées	260 (4.91%)	80 (5.01%)	180 (4.87%)	
Limousin	52 (0.98%)	19 (1.19%)	33 (0.89%)	
Rhône-Alpes	596 (11.26%)	171 (10.71%)	425 (11.5%)	
Auvergne	122 (2.3%)	44 (2.76%)	78 (2.11%)	
Languedoc-Roussillon	220 (4.16%)	70 (4.38%)	150 (4.06%)	
Provence-Alpes-Côte d'Azur	350 (6.61%)	112 (7.01%)	238 (6.44%)	
Corse	7 (0.13%)	2 (0.13%)	5 (0.14%)	0.077
Size Urban Area				
Small Municipality	1954 (36.92%)	548 (34.31%)	1406 (38.04%)	
Medium Municipality	554 (10.47%)	147 (9.2%)	407 (11.01%)	
Large Municipality	1587 (29.98%)	474 (29.68%)	1113 (30.11%)	

Continued on next page

Table A2 – Continued from previous page

Variable	Self-reported No depression	MHI-5 $\leq Q1$	MHI-5 $> Q1$	p-value
Paris metropolitan area	520 (9.82%)	143 (8.95%)	377 (10.2%)	
No answer	678 (12.81%)	285 (17.85%)	393 (10.63%)	$\mathbf{< 10^{-3}}$
Sunlight	2088.7 ($\pm$311.91)	2089.46 ($\pm$316.43)	2088.37 ($\pm$309.98)	0.907

Notes: The MHI-5 scores range from 0 to 100, with a score of 100 representing optimal mental health. $N = 5,293$. The first column gives the characteristics of the sample of people who self-reported not having depression regardless of the MHI-5 score. The second and third columns consider only those people whose MHI-5 scores is less than or equal to $Q1 = 60$ and strictly greater than $Q1 = 60$, respectively. The different characteristics are expressed in terms of the mean $\pm$ standard deviation for continuous variables and in terms of absolute and relative frequencies for nominal and ordinal variables. The fourth column provides a test p-value, depending on the type of variables (analysis of variance for continuous variables, χ^2-test for nominal variables, and Kruskal-Wallis rank sum test for ordinal variables). Bold values indicate significant differences (p-value < 0.05).

Table A3. Descriptive Statistics of Health Care Expenditures.

Variable	Self-reported No depression	MHI-5 $\leq Q1$	MHI-5 $> Q1$	p-value
Expenses				
Outpatient	1469.8 ($\pm$2528.39)	2024.47 ($\pm$3305.82)	1230.14 ($\pm$2060.32)	$\mathbf{< 10^{-3}}$
General Practitioner	125.64 ($\pm$143.02)	162.99 ($\pm$171.99)	109.5 ($\pm$125.12)	$\mathbf{< 10^{-3}}$
Specialist	193.62 ($\pm$376.86)	229.62 ($\pm$389.06)	178.07 ($\pm$370.44)	$\mathbf{< 10^{-3}}$
Pharmacy	468.79 ($\pm$1473.3)	684.48 ($\pm$1748.96)	375.6 ($\pm$1326.17)	$\mathbf{< 10^{-3}}$
Physiotherapist	59.12 ($\pm$221.51)	90.57 ($\pm$301.99)	45.53 ($\pm$173.97)	$\mathbf{< 10^{-3}}$
Nurse	70.77 ($\pm$620.48)	130.68 ($\pm$875.91)	44.89 ($\pm$466.68)	$\mathbf{< 10^{-3}}$
Dentist	164.77 ($\pm$503.46)	176.6 ($\pm$553.59)	159.66 ($\pm$480.18)	0.261
Equipment	80.91 ($\pm$499.86)	136.98 ($\pm$665.04)	56.68 ($\pm$406.04)	$\mathbf{< 10^{-3}}$
Transport	57.56 ($\pm$610.5)	121.82 ($\pm$1005.28)	29.8 ($\pm$307.78)	$\mathbf{< 10^{-3}}$
Optical	104.41 ($\pm$224.41)	107.36 ($\pm$222.11)	103.14 ($\pm$225.42)	0.53
Prostheses	39.27 ($\pm$248.88)	52.26 ($\pm$296.01)	33.66 ($\pm$225.3)	**0.013**
Emergency w/o hospitalization	14.96 ($\pm$51.1)	18.65 ($\pm$58.7)	13.36 ($\pm$47.36)	$\mathbf{< 10^{-3}}$
Reimbursement				
Outpatient	966.66 ($\pm$2322.13)	1445.77 ($\pm$3106.26)	759.64 ($\pm$1847.45)	$\mathbf{< 10^{-3}}$
General Practitioner	89.44 ($\pm$112.53)	120.49 ($\pm$141.45)	76.03 ($\pm$94.32)	$\mathbf{< 10^{-3}}$
Specialist	133.15 ($\pm$301.12)	163.41 ($\pm$333.25)	120.07 ($\pm$285.18)	$\mathbf{< 10^{-3}}$
Pharmacy	361.5 ($\pm$1431.61)	547.1 ($\pm$1688.87)	281.31 ($\pm$1296.91)	$\mathbf{< 10^{-3}}$
Physiotherapist	42.49 ($\pm$191.39)	68.65 ($\pm$273.58)	31.19 ($\pm$140.39)	$\mathbf{< 10^{-3}}$
Nurse	64.22 ($\pm$595.92)	123.79 ($\pm$870.91)	38.48 ($\pm$422.84)	$\mathbf{< 10^{-3}}$
Dentist	58.64 ($\pm$135.48)	63.06 ($\pm$156.12)	56.73 ($\pm$125.49)	0.119
Equipment	69.36 ($\pm$478.52)	121.45 ($\pm$640.45)	46.85 ($\pm$386.14)	$\mathbf{< 10^{-3}}$
Transport	54.8 ($\pm$606.35)	117.36 ($\pm$1000.33)	27.77 ($\pm$303.2)	$\mathbf{< 10^{-3}}$
Optical	2.63 ($\pm$6.17)	2.89 ($\pm$6.43)	2.52 ($\pm$6.05)	**0.046**
Prostheses	13.65 ($\pm$109.58)	18.65 ($\pm$136.82)	11.49 ($\pm$95.38)	**0.029**
Emergency w/o hospitalization	11.97 ($\pm$40.47)	15.03 ($\pm$46.65)	10.65 ($\pm$37.42)	$\mathbf{< 10^{-3}}$
Co-payment				
Outpatient	236.32 ($\pm$273.57)	286.12 ($\pm$312.7)	214.8 ($\pm$251.82)	$\mathbf{< 10^{-3}}$
General Practitioner	27.22 ($\pm$35.42)	31.73 ($\pm$40.59)	25.27 ($\pm$32.75)	$\mathbf{< 10^{-3}}$
Specialist	32.75 ($\pm$46.61)	37.31 ($\pm$52.35)	30.78 ($\pm$43.75)	$\mathbf{< 10^{-3}}$
Pharmacy	93.7 ($\pm$142.51)	120.43 ($\pm$176.69)	82.16 ($\pm$123.13)	$\mathbf{< 10^{-3}}$
Physiotherapist	14.73 ($\pm$52.04)	19.08 ($\pm$59.14)	12.85 ($\pm$48.54)	$\mathbf{< 10^{-3}}$
Nurse	5.59 ($\pm$79.13)	5.46 ($\pm$23.94)	5.65 ($\pm$93.39)	0.937
Dentist	21.56 ($\pm$52.58)	23.06 ($\pm$58.81)	20.91 ($\pm$49.64)	0.172
Equipment	9.56 ($\pm$58.31)	12.24 ($\pm$65.53)	8.4 ($\pm$54.87)	**0.028**
Transport	2.05 ($\pm$22.25)	3.8 ($\pm$34.09)	1.29 ($\pm$14.31)	$\mathbf{< 10^{-3}}$
Optical	1.57 ($\pm$3.81)	1.65 ($\pm$3.96)	1.53 ($\pm$3.75)	0.304
Prostheses	4.67 ($\pm$17.96)	5.16 ($\pm$18.63)	4.45 ($\pm$17.66)	0.192
Emergency w/o hospitalisation	2.73 ($\pm$11.51)	3.31 ($\pm$13.25)	2.47 ($\pm$10.66)	**0.015**
Extra-fees				
Outpatient	238.86 ($\pm$500.81)	257.66 ($\pm$540.78)	230.73 ($\pm$482.36)	0.073
General Practitioner	5.04 ($\pm$28.26)	5.89 ($\pm$31.36)	4.68 ($\pm$26.81)	0.151
Specialist	23.94 ($\pm$106.01)	24.48 ($\pm$86.38)	23.71 ($\pm$113.46)	0.808
Pharmacy	0.23 ($\pm$7.81)	0.28 ($\pm$7.74)	0.2 ($\pm$7.84)	0.72
Physiotherapist	0.81 ($\pm$13.9)	1.48 ($\pm$21.1)	0.52 ($\pm$9.18)	**0.02**
Nurse	0.02 ($\pm$0.44)	0.02 ($\pm$0.42)	0.02 ($\pm$0.45)	0.783
Dentist	84.57 ($\pm$344.5)	90.46 ($\pm$368.33)	82.02 ($\pm$333.7)	0.413
Equipment	1.99 ($\pm$33.16)	3.28 ($\pm$55.06)	1.43 ($\pm$16.25)	0.063
Transport	0.41 ($\pm$17.34)	0.06 ($\pm$2.04)	0.56 ($\pm$20.7)	0.328
Optical	100.21 ($\pm$217.26)	102.82 ($\pm$214.95)	99.09 ($\pm$218.27)	0.566

Continued on next page

Table A3 – Continued from previous page

Variable	**Self-reported No depression**	**MHI-5 $\leq Q1$**	**MHI-5 $> Q1$**	p-value
Prostheses	20.96 ($\pm$189.11)	28.46 ($\pm$222.58)	17.72 ($\pm$172.59)	0.058
Emergency w/o hospitalisation	0.02 ($\pm$0.51)	0.03 ($\pm$0.61)	0.02 ($\pm$0.47)	0.715
Deductible				
Outpatient	27.76 ($\pm$28.71)	34.7 ($\pm$32.96)	24.75 ($\pm$26.1)	$\mathbf{< 10^{-3}}$
General Practitioner	3.93 ($\pm$4.38)	4.88 ($\pm$5.24)	3.52 ($\pm$3.87)	$\mathbf{< 10^{-3}}$
Specialist	3.59 ($\pm$4.89)	4.24 ($\pm$5.48)	3.31 ($\pm$4.58)	$\mathbf{< 10^{-3}}$
Pharmacy	13.36 ($\pm$15.38)	16.66 ($\pm$17.08)	11.93 ($\pm$14.35)	$\mathbf{< 10^{-3}}$
Physiotherapist	1.09 ($\pm$3.54)	1.36 ($\pm$3.83)	0.98 ($\pm$3.41)	$\mathbf{< 10^{-3}}$
Nurse	0.94 ($\pm$4.13)	1.41 ($\pm$5.44)	0.74 ($\pm$3.39)	$\mathbf{< 10^{-3}}$
Dentist	0 ($\pm$0.11)	0.01 ($\pm$0.17)	0 ($\pm$0.06)	**0.01**
Transport	0.31 ($\pm$2.04)	0.6 ($\pm$3.13)	0.18 ($\pm$1.3)	$\mathbf{< 10^{-3}}$
Emergency w/o hospitalisation	0.24 ($\pm$0.92)	0.28 ($\pm$1.03)	0.22 ($\pm$0.87)	**0.025**
Medical Sessions				
No. Medical Sessions General Pract.	4.73 ($\pm$5.02)	6.11 ($\pm$6.23)	4.13 ($\pm$4.26)	$\mathbf{< 10^{-3}}$
No. Medical Sessions Specialist	3.44 ($\pm$4.44)	4.13 ($\pm$4.98)	3.14 ($\pm$4.15)	$\mathbf{< 10^{-3}}$

Notes: The MHI-5 scores range from 0 to 100, with a score of 100 representing optimal mental health. $N = 5,293$. The first column gives the characteristics for the sample of people who self-report not having depression regardless of the MHI-5 score. The second and third columns consider among this sample, only people whose MHI-5 score is less than or equal to $Q1 = 60$ and strictly greater than $Q1 = 60$, respectively. The different characteristics are expressed in terms of mean $\pm$ standard deviation for continuous variables, and in terms of absolute and relative frequencies for nominal and ordinal variables. The fourth column gives a test p-value, depending on variables' type (analysis of variance for continuous variables, χ^2-test for nominal variables and Kruskal-Wallis rank sum test for ordinal variables). Bold values indicate a significant differences (p-value < 0.05).

Table A4. Descriptive Statistics on Reasons for Waiver for Care.

Variable	**Self-reported No depression**	**MHI-5 $\leq Q1$**	**MHI-5 $> Q1$**	p-value
Expenses				
Waiver General Practitioner				
Yes	188 (3.55%)	110 (6.89%)	78 (2.11%)	
No	4024 (76.02%)	1201 (75.2%)	2823 (76.38%)	
No answer	1081 (20.42%)	286 (17.91%)	795 (21.51%)	$\mathbf{< 10^{-3}}$
Waiver Dental Care				
Yes	695 (13.13%)	311 (19.47%)	384 (10.39%)	
No	3517 (66.45%)	1000 (62.62%)	2517 (68.1%)	
No answer	1081 (20.42%)	286 (17.91%)	795 (21.51%)	$\mathbf{< 10^{-3}}$
Waiver Other Health Care				
Yes	169 (3.19%)	92 (5.76%)	77 (2.08%)	
No	4043 (76.38%)	1219 (76.33%)	2824 (76.41%)	
No answer	1081 (20.42%)	286 (17.91%)	795 (21.51%)	$\mathbf{< 10^{-3}}$
Waiver Health Care Too Far				
Yes	100 (1.89%)	63 (3.94%)	37 (1%)	
No	4112 (77.69%)	1248 (78.15%)	2864 (77.49%)	
No answer	1081 (20.42%)	286 (17.91%)	795 (21.51%)	$\mathbf{< 10^{-3}}$
Waiver Appointment Delay Too Long				
Yes	632 (11.94%)	264 (16.53%)	368 (9.96%)	
No	3580 (67.64%)	1047 (65.56%)	2533 (68.53%)	
No answer	1081 (20.42%)	286 (17.91%)	795 (21.51%)	$\mathbf{< 10^{-3}}$

Notes: The MHI-5 scores range from 0 to 100, with a score of 100 representing optimal mental health. $N = 5,293$. The first column gives the characteristics of the sample of people who self-reported not having depression regardless of the MHI-5 score. The second and third columns consider only those people whose MHI-5 scores is less than or equal to $Q1 = 60$ and strictly greater than $Q1 = 60$, respectively. The different characteristics are expressed in terms of the mean $\pm$ standard deviation for continuous variables and in terms of absolute and relative frequencies for nominal and ordinal variables. The fourth column provides a test p-value, depending on the type of variables (analysis of variance for continuous variables, χ^2-test for nominal variables, and Kruskal-Wallis rank sum test for ordinal variables). Bold values indicate significant differences (p-value < 0.05).

Table A5. Descriptive Statistics of Working Conditions.

Variable	**Self-reported No depression**	**MHI-5 $\leq Q1$**	**MHI-5 $> Q1$**	p-value
Expenses				
Have to Hurry to Do Job				
Always	391 (7.39%)	168 (10.52%)	223 (6.03%)	
Often	895 (16.91%)	290 (18.16%)	605 (16.37%)	

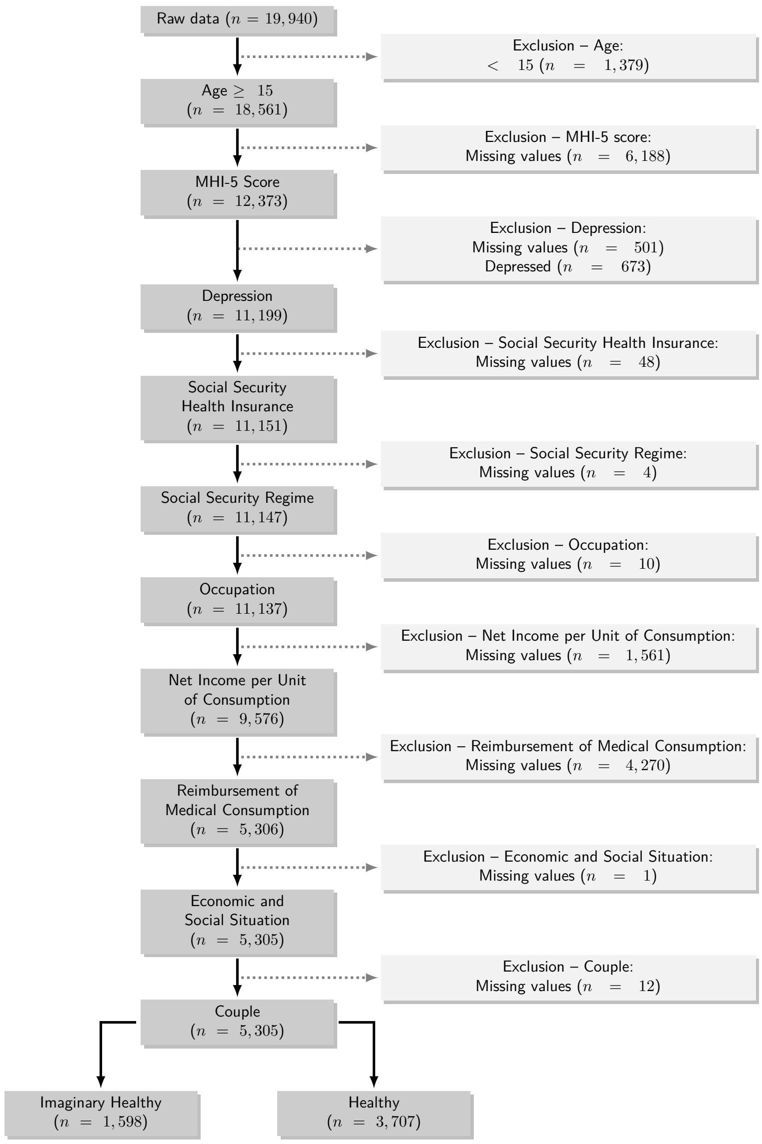
Continued on next page

Table A5 – Continued from previous page

Variable	Self-reported No depression	MHI-5 $\leq Q1$	MHI-5 $> Q1$	p-value
Sometimes	1016 (19.2%)	219 (13.71%)	797 (21.56%)	
Never	247 (4.67%)	39 (2.44%)	208 (5.63%)	
No answer	2744 (51.84%)	881 (55.17%)	1863 (50.41%)	$\mathbf{< 10^{-3}}$
Very Little Freedom to Do Job				
Always	126 (2.38%)	50 (3.13%)	76 (2.06%)	
Often	299 (5.65%)	130 (8.14%)	169 (4.57%)	
Sometimes	1084 (20.48%)	302 (18.91%)	782 (21.16%)	
Never	1028 (19.42%)	229 (14.34%)	799 (21.62%)	
No answer	2756 (52.07%)	886 (55.48%)	1870 (50.6%)	$\mathbf{< 10^{-3}}$
Job Allows to Learn New Things				
Always	401 (7.58%)	88 (5.51%)	313 (8.47%)	
Often	960 (18.14%)	255 (15.97%)	705 (19.07%)	
Sometimes	1039 (19.63%)	316 (19.79%)	723 (19.56%)	
Never	151 (2.85%)	60 (3.76%)	91 (2.46%)	
No answer	2742 (51.8%)	878 (54.98%)	1864 (50.43%)	$\mathbf{< 10^{-3}}$
Colleagues Help Carry out Tasks				
Always	291 (5.5%)	59 (3.69%)	232 (6.28%)	
Often	785 (14.83%)	204 (12.77%)	581 (15.72%)	
Sometimes	978 (18.48%)	302 (18.91%)	676 (18.29%)	
Never	241 (4.55%)	90 (5.64%)	151 (4.09%)	
Not concered	247 (4.67%)	60 (3.76%)	187 (5.06%)	
No answer	2751 (51.97%)	882 (55.23%)	1869 (50.57%)	$\mathbf{< 10^{-3}}$
Job Requires not to Sleep Betw. Midnight and 5 a.m.				
Always	61 (1.15%)	18 (1.13%)	43 (1.16%)	
Often	91 (1.72%)	20 (1.25%)	71 (1.92%)	
Sometimes	254 (4.8%)	77 (4.82%)	177 (4.79%)	
Never	2136 (40.36%)	605 (37.88%)	1531 (41.42%)	
No answer	2751 (51.97%)	877 (54.92%)	1874 (50.7%)	**0.037**
Repetitive Work under Time Constraints / Line Job				
Always	174 (3.29%)	64 (4.01%)	110 (2.98%)	
Often	227 (4.29%)	79 (4.95%)	148 (4%)	
Sometimes	330 (6.23%)	100 (6.26%)	230 (6.22%)	
Never	1809 (34.18%)	474 (29.68%)	1335 (36.12%)	
No answer	2753 (52.01%)	880 (55.1%)	1873 (50.68%)	$\mathbf{< 10^{-3}}$
Exposed to Carrying Heavy Loads				
Always	187 (3.53%)	68 (4.26%)	119 (3.22%)	
Often	392 (7.41%)	127 (7.95%)	265 (7.17%)	
Sometimes	725 (13.7%)	200 (12.52%)	525 (14.2%)	
Never	1237 (23.37%)	323 (20.23%)	914 (24.73%)	
No answer	2752 (51.99%)	879 (55.04%)	1873 (50.68%)	$\mathbf{< 10^{-3}}$
Exposed to Painful Postures				
Always	409 (7.73%)	153 (9.58%)	256 (6.93%)	
Often	523 (9.88%)	148 (9.27%)	375 (10.15%)	
Sometimes	533 (10.07%)	125 (7.83%)	408 (11.04%)	
Never	1078 (20.37%)	294 (18.41%)	784 (21.21%)	
No answer	2750 (51.96%)	877 (54.92%)	1873 (50.68%)	$\mathbf{< 10^{-3}}$
Exposed to Harmful/Toxic Products/Substances				
Always	211 (3.99%)	77 (4.82%)	134 (3.63%)	
Often	335 (6.33%)	92 (5.76%)	243 (6.57%)	
Sometimes	582 (11%)	150 (9.39%)	432 (11.69%)	
Never	1415 (26.73%)	399 (24.98%)	1016 (27.49%)	
No answer	2750 (51.96%)	879 (55.04%)	1871 (50.62%)	**0.002**

Notes: The MHI-5 scores range from 0 to 100, with a score of 100 representing optimal mental health. $N = 5,293$. The first column gives the characteristics of the sample of people who self-reported not having depression regardless of the MHI-5 score. The second and third columns consider only those people whose MHI-5 scores is less than or equal to $Q1 = 60$ and strictly greater than $Q1 = 60$, respectively. The different characteristics are expressed in terms of the mean $\pm$ standard deviation for continuous variables and in terms of absolute and relative frequencies for nominal and ordinal variables. The fourth column provides a test p-value, depending on the type of variables (analysis of variance for continuous variables, χ^2-test for nominal variables, and Kruskal-Wallis rank sum test for ordinal variables). Bold values indicate significant differences (p-value < 0.05).

Figure A1. Successive Steps in the Selection of Individuals in the Final Sample.



Notes: The red boxes show the criteria used to filter the data, the reason for deletion, and the number of individuals removed from the sample. The green boxes indicate the number of observations remaining after the removal of observations. For example, the first green cell indicates that the initial sample consists of $19,940$ observations. The first red cell indicates that $1,379$ individuals in the initial sample are younger than 15 years of age. These individuals are removed, leaving $18,561$ observations, as indicated by the second green cell.

Table A6. Descriptive Statistics of Economic and Social Situation.

Variable	Self-reported No depression	MHI-5 $\leq Q1$	MHI-5 $> Q1$	p-value
Expenses				
Participation in Group Activities				
Yes	1848 (34.91%)	458 (28.68%)	1390 (37.61%)	
No	3337 (63.05%)	1103 (69.07%)	2234 (60.44%)	
No answer	108 (2.04%)	36 (2.25%)	72 (1.95%)	$< 10^{-3}$
Frequency Meeting with Friends/Neighbors				
Every day or almost every day	1144 (21.61%)	300 (18.79%)	844 (22.84%)	
At least once a week	2150 (40.62%)	616 (38.57%)	1534 (41.5%)	
At least once a month	1098 (20.74%)	325 (20.35%)	773 (20.91%)	

Continued on next page

Table A6 – Continued from previous page

Variable	Self-reported No depression	MHI-5 $\leq Q1$	MHI-5 $> Q1$	p-value
Less than once a month	497 (9.39%)	181 (11.33%)	316 (8.55%)	
Never	270 (5.1%)	129 (8.08%)	141 (3.81%)	
No answer	134 (2.53%)	46 (2.88%)	88 (2.38%)	$\mathbf{< 10^{-3}}$
Frequency Meeting with People in Organizations				
Every day or almost every day	161 (3.04%)	49 (3.07%)	112 (3.03%)	
At least once a week	912 (17.23%)	209 (13.09%)	703 (19.02%)	
At least once a month	604 (11.41%)	156 (9.77%)	448 (12.12%)	
Less than once a month	657 (12.41%)	178 (11.15%)	479 (12.96%)	
Never	2717 (51.33%)	923 (57.8%)	1794 (48.54%)	
No answer	242 (4.57%)	82 (5.13%)	160 (4.33%)	$\mathbf{< 10^{-3}}$
Frequency Meeting with Colleagues Outside Work				
Every day or almost every day	275 (5.2%)	68 (4.26%)	207 (5.6%)	
At least once a week	446 (8.43%)	128 (8.02%)	318 (8.6%)	
At least once a month	670 (12.66%)	175 (10.96%)	495 (13.39%)	
Less than once a month	1032 (19.5%)	275 (17.22%)	757 (20.48%)	
Never	2374 (44.85%)	808 (50.59%)	1566 (42.37%)	
No answer	496 (9.37%)	143 (8.95%)	353 (9.55%)	$\mathbf{< 10^{-3}}$
Frequency Meeting with Family Living Outside Household				
Every day or almost every day	1022 (19.31%)	311 (19.47%)	711 (19.24%)	
At least once a week	1969 (37.2%)	553 (34.63%)	1416 (38.31%)	
At least once a month	1157 (21.86%)	323 (20.23%)	834 (22.56%)	
Less than once a month	783 (14.79%)	251 (15.72%)	532 (14.39%)	
Never	208 (3.93%)	93 (5.82%)	115 (3.11%)	
No answer	154 (2.91%)	66 (4.13%)	88 (2.38%)	$\mathbf{< 10^{-3}}$
Social Background				
Mother's Level of Education				
Never Been to School	355 (6.71%)	141 (8.83%)	214 (5.79%)	
Nursery School, Primary school, Certificate of studies	1996 (37.71%)	603 (37.76%)	1393 (37.69%)	
1st cycle (Middle School)	968 (18.29%)	266 (16.66%)	702 (18.99%)	
2nd cycle (High School	458 (8.65%)	107 (6.7%)	351 (9.5%)	
Higher Education	478 (9.03%)	134 (8.39%)	344 (9.31%)	
Other	82 (1.55%)	16 (1%)	66 (1.79%)	
Does not know	857 (16.19%)	300 (18.79%)	557 (15.07%)	
No answer	99 (1.87%)	30 (1.88%)	69 (1.87%)	$\mathbf{< 10^{-3}}$
Father's Level of Education				
Never Been to School	292 (5.52%)	117 (7.33%)	175 (4.73%)	
Nursery School, Primary school, Certificate of studies	1830 (34.57%)	557 (34.88%)	1273 (34.44%)	
1st cycle (Middle School)	1037 (19.59%)	269 (16.84%)	768 (20.78%)	
2nd cycle (High School	401 (7.58%)	99 (6.2%)	302 (8.17%)	
Higher Education	565 (10.67%)	154 (9.64%)	411 (11.12%)	
Other	123 (2.32%)	44 (2.76%)	79 (2.14%)	
Does not know	948 (17.91%)	319 (19.97%)	629 (17.02%)	
No answer	97 (1.83%)	38 (2.38%)	59 (1.6%)	$\mathbf{< 10^{-3}}$

Notes: The MHI-5 scores range from 0 to 100, with a score of 100 representing optimal mental health. $N = 5,293$. The first column gives the characteristics of the sample of people who self-reported not having depression regardless of the MHI-5 score. The second and third columns consider only those people whose MHI-5 scores is less than or equal to $Q1 = 60$ and strictly greater than $Q1 = 60$, respectively. The different characteristics are expressed in terms of the mean $\pm$ standard deviation for continuous variables and in terms of absolute and relative frequencies for nominal and ordinal variables. The fourth column provides a test p-value, depending on the type of variables (analysis of variance for continuous variables, χ^2-test for nominal variables, and Kruskal-Wallis rank sum test for ordinal variables). Bold values indicate significant differences (p-value < 0.05).

Table A7. Comparison of Respondents Having not Reported Mental Health Issues and Respondents Having not Answered

Variable	Self-reported No depression	MHI-5 $\leq Q1$	MHI-5 $> Q1$	p-value
Expenses				
MHI-5 Score (n = 10,398)	71.12 ($\pm$17.05)	71.13 ($\pm$17.05)	70.85 ($\pm$17.18)	0.73
Self-Assessed Health Condition (n = 11,063)				
Very Good	2661 (24.05%)	2511 (23.96%)	150 (25.82%)	

Continued on next page

Table A7 – Continued from previous page

Variable	Self-reported No depression	MHI-5 $\leq Q1$	MHI-5 $> Q1$	p-value
Good	5194 (46.95%)	4934 (47.07%)	260 (44.75%)	
Fairly Good	2637 (23.84%)	2493 (23.78%)	144 (24.78%)	
Bad	507 (4.58%)	483 (4.61%)	24 (4.13%)	
Very Bad	64 (0.58%)	61 (0.58%)	3 (0.52%)	0.74
Age (n = 15,699)	46.64 ($\pm$18.8)	48.26 ($\pm$19.1)	43.3 ($\pm$17.71)	$< 10^{-3}$
Gender (n = 15,699)				
Female	7889 (50.25%)	5411 (51.25%)	2478 (48.2%)	
Male	7810 (49.75%)	5147 (48.75%)	2663 (51.8%)	$< 10^{-3}$
Occupation (n = 15,674)				
Farmer	575 (3.67%)	413 (3.91%)	162 (3.16%)	
Craftsman, trader	838 (5.35%)	555 (5.26%)	283 (5.52%)	
Executive and intellectual profession	1985 (12.66%)	1437 (13.62%)	548 (10.69%)	
Intermediate occupation	2649 (16.9%)	1866 (17.69%)	783 (15.28%)	
Administrative employee	2097 (13.38%)	1466 (13.9%)	631 (12.31%)	
Commercial employee	1773 (11.31%)	1187 (11.25%)	586 (11.44%)	
Skilled worker	2478 (15.81%)	1578 (14.96%)	900 (17.56%)	
Unskilled worker	1133 (7.23%)	718 (6.81%)	415 (8.1%)	
Inactive having never worked	2146 (13.69%)	1330 (12.61%)	816 (15.93%)	$< 10^{-3}$
Professional Status (n = 13,564)				
Public employee	2710 (19.98%)	1880 (20.35%)	830 (19.2%)	
Private employee	8613 (63.5%)	5833 (63.13%)	2780 (64.29%)	
Other	2241 (16.52%)	1527 (16.53%)	714 (16.51%)	0.27
Social Security (n = 15,652)				
Yes, because he/she works, or receives unemployment benefits, is a student, retired, widower of a pensioner, disabled, beneficiary of the basic CMU...	13707 (87.57%)	9308 (88.4%)	4399 (85.87%)	
Yes, as the beneficiary of a person living in your household	1945 (12.43%)	1221 (11.6%)	724 (14.13%)	$< 10^{-3}$
Social Security System (n = 15,682)				
The general scheme (Cnamts)	11142 (71.05%)	7477 (70.89%)	3665 (71.37%)	
Public service	1401 (8.93%)	943 (8.94%)	458 (8.92%)	
The local Alsace-Moselle scheme	506 (3.23%)	359 (3.4%)	147 (2.86%)	
The basic Universal Health Coverage	32 (0.2%)	19 (0.18%)	13 (0.25%)	
The agricultural scheme	1075 (6.85%)	767 (7.27%)	308 (6%)	
The self-employed scheme	1008 (6.43%)	653 (6.19%)	355 (6.91%)	
Other (Student, abroad, other)	493 (3.14%)	310 (2.94%)	183 (3.56%)	
Does not know	25 (0.16%)	19 (0.18%)	6 (0.12%)	$< 10^{-3}$
Couple (n = 14,236)_ *Yes*	10951 (76.92%)	7524 (79.43%)	3427 (71.95%)	
No	3285 (23.08%)	1949 (20.57%)	1336 (28.05%)	$< 10^{-3}$
Marital Status (n = 15,401)				
Married, Civil Union	9142 (59.36%)	6327 (60.94%)	2815 (56.1%)	
Divorced, separated	588 (3.82%)	419 (4.04%)	169 (3.37%)	
Widowed	597 (3.88%)	469 (4.52%)	128 (2.55%)	
Lives in a marriage or concubinage	1803 (11.71%)	1192 (11.48%)	611 (12.18%)	
Single	3253 (21.12%)	1965 (18.93%)	1288 (25.67%)	
Does not know	6 (0.04%)	4 (0.04%)	2 (0.04%)	
Refuses to answer	12 (0.08%)	7 (0.07%)	5 (0.1%)	$< 10^{-3}$
Long-term condition (Self-declared) (n = 15,699)				
Yes	2529 (16.11%)	1877 (17.78%)	652 (12.68%)	
No	13059 (83.18%)	8651 (81.94%)	4408 (85.74%)	
Does not know	111 (0.71%)	30 (0.28%)	81 (1.58%)	$< 10^{-2}$

Notes: The MHI-5 scores range from 0 to 100, with a score of 100 representing optimal mental health. $N = 5,293$. The first column gives the characteristics of the sample of people who self-reported not having depression regardless of the MHI-5 score. The second and third columns consider only those people whose MHI-5 scores is less than or equal to $Q1 = 60$ and strictly greater than $Q1 = 60$, respectively. The different characteristics are expressed in terms of the mean $\pm$ standard deviation for continuous variables and in terms of absolute and relative frequencies for nominal and ordinal variables. The fourth column provides a test p-value, depending on the type of variables (analysis of variance for continuous variables, χ^2-test for nominal variables, and Kruskal-Wallis rank sum test for ordinal variables). Bold values indicate significant differences (p-value < 0.05).

B Classifiers

We train four models to predict the *imaginary healthy* patients: random forest, extreme gradient boosting, support vector machine, and penalised logistic regression. This section briefly describes the models and details the estimation procedure, followed by an overview of their performance using the AUC metric.

B.1 Algorithms

Random Forests. RF (Breiman, 2001) are an ensemble learning method that operates by constructing multiple decision trees. The algorithm uses a bagging technique (bootstrap aggregation) during the training process, which involves randomly selecting observations to train each tree and randomly selecting variables when performing space partitioning at each node of the trees. For classification tasks, each terminal node of the individual trees outputs the class that is the mode of the classes. The output of the random forest is determined by the majority vote of the outputs from the individual trees. To associate a level of confidence with the prediction, it is common practice to calculate the proportion of positive class observations among all predictions made by the ensemble of trees; this value is referred to as a prediction score. RF are convenient for applications in high-dimensional data; their use is recommended in health outcome research (Crown, 2015), and they have been recently used to detect depression (Pearson et al., 2019; Cacheda et al., 2019). We train the forests using the `ranger` package in R. The Gini index is employed as the splitting rule, and the number of trees is set to 500. In the grid search process, we vary two hyperparameters: the minimum size of a tree node (with values of 50, 75, 100, and 150) and the number of variables considered for a split, ranging from 3 to the square root of the total number of explanatory variables.

Extreme Gradient Boosting. The gradient boosting algorithm (Chen and Guestrin, 2016), implemented in XGBoost, sequentially builds a series of decision trees where each new tree corrects the errors made by the previous ones. At each iteration, the residuals from the previous tree are estimated, and a regularisation term is used to prevent overfitting. We train the models using the `xgboost` R package. We set the number of boosting iterations to 500 and the learning rate to 0.01. Several hyperparameters are varied. The maximum depth of the trees is varied within the range $[3, 6]$. Deeper trees are more complex and present a higher risk of overfitting. We also vary the fraction of columns to be randomly sampled for each tree, choosing values from the set $\{0.1, 0.2, \ldots, 1\}$. This hyperparameter acts as a form of regularisation, helping to mitigate overfitting and improve generalisation by generating more diverse trees. Additionally, we vary the hyperparameter that specifically controls regularisation, considering the possible values of 0, 5, and 10. When this hyperparameter is set to 0, tree splits are made as long as they provide a gain. Increasing this value leads to pruning of the trees, which can again help prevent overfitting. Finally, we explore different values for the proportion of training data used to construct the trees: 70%, 80%, 90%, and the entire training sample.

Support Vector Machine. SVM (Ben-Hur et al., 2002) finds the hyperplane that best separates the observations of different classes. In our binary classification case, the best hyperplane maximises the distance—called the margin—between the hyperplane and the nearest data points from each class—called the support vectors. To handle non-linear relationships between data points, the input data is transformed into a higher-dimensional space where a linear separation is possible using a radial kernel. More specifically the kernel writes $K(x, x') = \exp(-\gamma \|x - x'\|_2^2)$, where x and x' are two data points, $\|x - x'\|_2^2$ is the squared Euclidean distance between them, and γ is a parameter that determines the kernel's width. We train the model using the `e1071` R library. The training data is centered and scaled. The value for γ is estimated using the `sigest` function from the `kernlab` R package. We let the regularisation parameter, which controls the trade-off between maximising the margin and minimising the classification error, vary in values in the set made of 2 to the power of each element in the sequence $\{-2, -1, 0, 1, \ldots, 97\}$.[15] The lower this cost hyperparameter, the more the model will allow for some misclassifications. Conversely, the higher this hyperparameter, the more complex the model, which may result in overfitting.

Penalised Logistic Regression Model. Penalised logistic regression (Friedman et al., 2010) is an extension of logistic regression that incorporates regularisation. Regularisation is achieved by adding a penalty term to the logistic regression's objective function, which constrains the size of the coefficients in the following problem:

$$\min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} l(y_i, \beta_0 + \beta^T x_i) + \underbrace{\lambda \left[(1-\alpha)\|\beta\|_2^2/2 + \alpha\|\beta\|_1\right]}_{\text{elastic net penalty}},$$

where N is the total number of observations, y_i is the binary target value, x_i is the vector of predictor values, $l(y_i, \beta_0 + \beta^T x_i)$ is the negative log-likelihood contribution for observation i. The penalty term is a weighted combination of the L1 norm (lasso) and the L2 norm (ridge), controlled by the elastic net mixing parameter, α. When $\alpha = 1$ (respectively $\alpha = 0$), the penalty is purely lasso (respectively ridge). The intercept β_0 and the coefficients β are to be estimated. When training the classifier, we vary values of α over a grid (sequence of 100 values evenly spaced from 0.1 to 1) to find the optimal balance between the L1 and L2 penalties. For each value of α, the regularisation parameter λ is determined through cross-validation so as to maximise the sensitivity on the validation set. This cross validation is performed using the `glmnet` function from the `glmnet` R package.

B.2 Grid Search

For each algorithm, the choice of hyperparameters is made using grid search, with repeated k-fold cross-validation (10 repetitions, 5 folds). The possible values for the hyperparameters are listed in Table B8. The training process is performed using the `caret` R package. See the online replication ebook for more details on the R codes.

After the grid search, the hyperparameters for the models are the following:

- Random Forest: `mtry=9`, `min.node.size=50`
- XGBoost: `nrounds=500`, `max_depth=5`, `colsample_bytree=8`, `eta=0.01`, `gamma=5`, `min_child_weight=150`, `subsample=0.9`
- SVM: `gamma=0.003586236`, `cost=0.25`
- Penalized Logistic Regression: `alpha=0.2909091`, `lambda=0.01157483`.

B.3 Model Performances

The ROC curve obtained from the train and the test sets on the four different models (random forest, XGBoost, SVM, and penalised logistic regression) are shown in Figure B2. The ROC curve represents the variation in the performance in terms of the true positive rate (TPR) and false positive rate (FPR) induced by varying the threshold above which an individual is predicted to be positive (*imaginary healthy* patient). The area under the curve (ROC-AUC) on the test set is about 0.7 for all models (see Table 2). If the model were able to perfectly predict the class of respondents, this area would be 1. Conversely, if the model could not correctly predict any individual, this area would be 0.

[15] More specifically, $2^{(3:(\text{len}+2))-5}$ where len equals 100.

Table B8. Hyperparameters.

Hyperparameter	Possible Values	Description
	Random Forest	
`mtry`	{3, 4, 5, 6, 7, 8, **9**}	Number of variables randomly sampled as candidates at each split
`splitrule`	Gini index	Splitting rule
`min.node.size`	{50, 75, 100, 150}	Minimum size of terminal nodes
	Extreme Gradient Boosting	
`nrounds`	500	Number of boosting iterations
`max_depth`	{3, 4, 5, 6}	Maximum depth of a tree
`colsample_bytree`	{.1, .2, ..., .9}	Subsample ratio of columns when constructing each tree
`eta`	0.01	Learning rate
`gamma`	{0, 5, 10}	Minimum loss reduction required to make a further partition on a leaf node of the tree
`min_child_weight`	{50, 100, 150}	Minimum sum of instance weight needed in a child
`subsample`	{0.7, 0.8, 0.9, 1}	Subsample ratio of the training instances
	Support Vector Machine (with radial kernel)	
`gamma`	Estimated with the `sigest()` function from the `kernlab` package	Radial kernel parameter
`cost`	$\{2^{-2}, 2^{-1}, \ldots, 2^{97}\}$	Cost of constraints violation
	Penalised Logistic Regression Model.	
`alpha`	Sequence of equally distant values from 0.1 to 1 with a length of 100	Elasticnet mixing parameter

Figure B2. ROC Curve on Validation Data.