Comprehensive Psychiatry 123 (2023) 152380

Contents lists available at ScienceDirect

Comprehensive Psychiatry

journal homepage: www.elsevier.com/locate/comppsych

Identifying populations at ultra-high risk of suicide using a novel machine learning method

Guus Berkelmans [a,b,c,*], Lizanne Schweren [b], Sandjai Bhulai [c], Rob van der Mei [a,c], Renske Gilissen [b]

[a] Centrum Wiskunde & Informatica, Science Park 123, 1098XG Amsterdam, The Netherlands
[b] 113 zelfmoordpreventie, Paasheuvelweg 25, 1105BP Amsterdam, The Netherlands
[c] Vrije Universiteit Amsterdam, De Boelelaan 1111, 1081HV Amsterdam, The Netherlands

ARTICLE INFO

Keywords:
Suicide
Risk factors
Population data
Machine learning
Interactions

ABSTRACT

Background: Targeted interventions for suicide prevention rely on adequate identification of groups at elevated risk. Several risk factors for suicide are known, but little is known about the interactions between risk factors. Interactions between risk factors may aid in detecting more specific sub-populations at higher risk.
Methods: Here, we use a novel machine learning heuristic to detect sub-populations at ultra high-risk for suicide based on interacting risk factors. The data-driven and hypothesis-free model is applied to investigate data covering the entire population of the Netherlands.
Findings: We found three sub-populations with extremely high suicide rates (i.e. >50 suicides per 100,000 person years, compared to 12/100,000 in the general population), namely: (1) people on unfit for work benefits that were never married, (2) males on unfit for work benefits, and (3) those aged 55–69 who live alone, were never married and have a relatively low household income. Additionally, we found two sub-populations where the rate was higher than expected based on individual risk factors alone: widowed males, and people aged 25–39 with a low level of education.
Interpretation: Our model is effective at finding ultra-high risk groups which can be targeted using sub-population level interventions. Additionally, it is effective at identifying high-risk groups that would not be considered risk groups based on conventional risk factor analysis.

1. Introduction

In the Netherlands alone, an average of five people die by suicide each day [1]. Every case of suicide marks a personal tragedy, both for the victim and for those left behind. Therefore, it is of utmost importance to implement effective suicide prevention programmes at multiple levels, including interventions directed at the entire population (e.g., public awareness campaigns), interventions targeting high-risk groups or sub-populations (e.g., training gatekeepers among professionals encountering individuals with financial difficulties) and interventions targeting at-risk individuals (e.g., cognitive behavioural therapy for individuals with suicidal thoughts) [2].

Interventions at the second level, targeting sub-populations, require adequate identification and detection of groups at elevated risk of suicide. Multiple studies have been performed to detect risk factors for suicide [3–7]. Not unexpectedly, the most important predictor of death by suicide is a prior non-fatal suicide attempt or prior psychiatric hospitalization [6]. Experiencing stressful life events and mental health problems including depression and substance use problems substantially increase the risk for suicide attempts and suicidal ideation, which in turn increases the risk of suicide [6]. In addition, certain socio-demographic groups are at elevated risk, including but not limited to men, people of middle age, people of lower socio-economic status and people living alone [6,1].

In complex and multifactorial outcomes such as mental illness, risk factors are known to interact or accumulate. For instance, stressful life events may trigger a depressive episode in persons with a genetic vulnerability to depression [8]. To our knowledge, however, little is known about interacting socio-demographic risk factors for suicide. In a hypothetical example, one might expect that unemployment might increase the risk of suicide more for men living alone than for the rest of the population. The detection of relevant interacting socio-demographic

* Corresponding author at: Centrum Wiskunde & Informatica, Science Park 123, 1098XG Amsterdam, The Netherlands.
E-mail address: g.a.berkelmans@vu.nl (G. Berkelmans).

https://doi.org/10.1016/j.comppsych.2023.152380
Received 4 August 2022; Accepted 24 February 2023
Available online 1 March 2023

Table 1
Predictor variables or 'features of interest' included in the machine learning model, after sampling (all person years resulting in suicide were included and 3% of the person years not resulting in suicide were included, see model section), (ref) means the reference category.

Features	Response categories	N	%
Sex	Male (ref)	2050131	49.4
	Female	2097177	50.6
Age in years	10–24	835473	20.1
	25–39 (ref)	856591	20.7
	40–54	999010	24.1
	55–69	879303	21.2
	70+	576931	13.9
Immigration background	Dutch (ref)	3231078	77.9
	1st generation western	213524	5.1
	2nd generation western	207883	5.0
	1st generation non-western	314951	7.6
	2nd generation non-western	179868	4.3
Personal income	1st quartile (ref)	1007657	24.3
	2nd quartile	1027422	24.8
	3rd quartile	1016962	24.5
	4th quartile	1015324	24.5
	Unknown	79943	1.9
Household income	1st quartile (ref)	1019868	24.6
	2nd quartile	1016622	24.5
	3rd quartile	1016383	24.5
	4th quartile	1014626	24.5
Household wealth/debts	1st quartile (ref)	1017399	24.5
	2nd quartile	1017837	24.5
	3rd quartile	1016503	24.5
	4th quartile	1015760	24.5
Level of education	Low	892702	21.5
	Middle (ref)	859185	20.7
	High	684749	16.5
	Unknown	1710672	41.3
Physical healthcare costs	€0 (ref)	59793	1.4
	€1–€5000,	3635734	87.7
	€5001–€10000	201167	4.9
	€10001+	183200	4.4
	Unknown	67414	1.6
Place in household	Child living at home	760069	18.3
	Living alone	802714	19.4
	Partner in couple with children	1201518	29.0
	Partner couple without children (ref)	1102279	26.6
	Other	280728	6.8
Marital status	Never married/registered partner (ref)	1714362	41.3
	Married/registered partner	1834896	44.3
	Divorced	348547	8.4
	Widowed	232123	5.6
Unfit for work benefits	Yes	196522	4.7
	No (ref)	3950786	95.3
Short-term unemployment benefits	Yes	215734	5.2
	No (ref)	3931574	94.8
Long-term unemployment benefits	Yes	171810	4.1
	No (ref)	3975498	95.9

risk factors will allow the identification of more specific sub-populations at elevated risk of suicide. This may increase the efficacy of targeted preventive interventions and has the potential to reduce suicide rates.

Machine learning methods offer new possibilities for flexible, data-driven, hypothesis-free and robust investigation of accumulating risk factors for suicide. A recent study performed such analyses using predominantly healthcare data and succeeded in identifying multiple relevant interactions [9]. Risk of suicide was higher, for instance, in men and women who had recently attempted suicide and were not being treated with pharmacotherapy. In a second study, including over 15,000 features (including but not limited to: demographics, diagnostic codes, procedure codes, and medication prescriptions) in the initial model and retaining 117 of them, researchers were able to develop a risk prediction model with acceptable performance parameters to stratify hospital patients by suicide risk [10].

An important limitation of the above studies is their complexity, hampering translation of their results to actionable recommendations for clinical practice. Moreover, as Kirtley et al. have recently emphasized [11], current machine learning methods have limited capabilities to support decisions and interventions at the individual level, as false-positive rates as well as false-negative rates are typically high. Thus, there is a need for more actionable and transparent machine-learning models to aid detection of high-risk subgroups rather than individuals.

In this paper, we present a new machine learning model that allows for investigation of complex interactions of socio-demographic risk factors whilst retaining interpretability. This model is applied to predict suicide risk groups in a dataset spanning the entire population of the Netherlands over a period of nine years, thereby mitigating sampling bias and sample size limitations. Our model yields detailed and interpretable results to aid the identification of sub-populations of individuals at relatively high risk for suicide, which may aid targeted preventive interventions.

2. Material and methods

2.1. Data

Statistic Netherlands (CBS) is a national administrative authority aiming to collect and provide reliable information that advances the understanding of social issues. CBS maintains a high-quality database containing, among others, socio-demographic and medical information regarding every inhabitant of the Netherlands. Analyses on CBS data are to be performed via a remote access connection to their computational servers. All results are verified prior to release, ensuring compliance with privacy laws.

For the current paper, we included data regarding all inhabitants of the Netherlands on the 31st of December of nine consecutive years (2011 to 2019), adding up to a total of 137,666,515 person years. Of those, 16,417 person years ended by suicide in the year following observation and 137,650,098 person years did not end by suicide in the year following observation.

2.2. Features of interest

The following socio-demographic predictor variables were measured on the 31st of December of the year preceding the outcome: sex, age, immigration background, household income, personal income, household wealth or debts, level of education, physical healthcare costs, place in household, marital status, short-term unemployment benefits, long-term unemployment benefits and unfit for work benefits. For details, see Table 1. Categorical variables were one-hot-encoded for use in machine learning analyses, meaning that for each category a new variable was introduced which has value 1 if the individual was in said category and has value 0 otherwise. Continuous variables were split into mutually exclusive response categories (e.g., quartiles) and also one-hot-encoded.

2.3. Model

A heuristic algorithm was devised to obtain interacting features which provide additional risk of suicide or reduce the risk. The obtained interaction features were prioritised on statistical significance as well as model improvement. The algorithm comprises four steps.

Step 1: the data is divided into three disjoint partitions: a training set, a validation set and a test set. The training set includes fifty percent of person years ending in suicide (N = 8,214) and one percent of all other person years (N = 1,377,055) and is used to detect significant interactions between features of interest. The validation set includes forty percent of person years ending in suicide (N = 6,512) and one percent of all other person years (N = 1,377,870) and is used to estimate the final logistic regression model. The test set includes ten percent of

Table 2
Interaction terms found by the algorithm as tested on the validation set. With corresponding Beta parameters, Odds-Ratios, Compound Odds Ratios, absolute and relative number of suicides within the sub-population within the validation set. Sub-populations with ⩾30 suicides per 100,000 are in bold.

Interaction term	Beta (95% CI)	Odds-Ratio (95% CI)	Compound Odds Ratio (95%CI)	Number of suicides	Relative number of suicides
Aged 25–39 and low level of education	0.46 ([0.30, 0.62])	1.58 (1.35, 1.86])	1.63 ([1.38, 1.93])	259	20.07
Aged 40–54 and long-term unemployment	**−0.22 ([−0.41, −0.04])**	**0.80 ([0.67, 0.96])**	**2.23 ([1.90, 2.61])**	**234**	**35.58**
Aged 55–69 and living alone	**−0.42 ([−0.67,−0.17])**	**0.66 ([0.51, 0.84])**	**2.27 ([1.78, 2.9])**	**833**	**35.54**
Aged 55–69 and living alone and Dutch immigration background	0.18 ([−0.04, 0.39])	1.20 ([0.96, 1.48])	2.71 ([2.30, 3.19])	728	39.37
Aged 55–69 and living alone and household income in the 1st quartile and never married	**−0.21 ([−0,43, 0.01])**	**0.81 ([0.65, 1.01])**	**3.44 ([2.60, 4.55])**	**229**	**57.22**
Aged 55–69 and never married	**0.32 ([0.15, 0.5])**	**1.38 ([1.16, 1.65])**	**2.00 ([1.64, 2.44])**	**427**	**34.81**
Aged 55–69 and part of couple without child at home	−0.46 ([−0.63, −0.29])	0.63 ([0.53, 0.75])	0.91 ([0.79, 1.05])	622	9.38
Aged 55–69 and healthcare costs of €10001 or more	**−0.44 ([−0.63, −0.25])**	**0.64 ([0.53, 0.78])**	**4.30 ([3.16, 5.86])**	**238**	**30.70**
Aged 70 or older and healthcare costs of €10001 or more	−0.66 ([−0.88, −0.44])	0.52 ([0.41, 0.64])	2.14 ([1.58, 2.9])	175	15.59
Male and unfit for work	**−0.39 ([−0.54, −0.24])**	**0.68 ([0.59, 0.78])**	**2.48 ([2.21, 2.79])**	**642**	**58.56**
Male and part of couple with child at home	0.64 ([0.48, 0.8])	1.90 ([1.61, 2.22])	0.82 ([0.73, 0.92])	801	10.94
Male and widowed	**0.54 ([0.33, 0.74])**	**1.72 ([1.40, 2.09])**	**1.56 ([1.31, 1.86])**	**218**	**31.31**
Male and healthcare costs of €10001 or more	−0.30 ([−0.46, −0.14])	0.74 ([0.63, 0.87])	3.42 ([2.64, 4.43])	456	27.48
Never married and unfit for work	**−0.03 ([−0.26, 0.19])**	**0.97 ([0.77, 1.21])**	**3.54 ([2.77, 4.53])**	**441**	**88.48**
Never married and unfit for work and physical healthcare costs between €1 and €5000	**0.54 ([0.31, 0.78])**	**1.72 ([1.36, 2.18])**	**6.45 ([4.83, 8.61])**	**321**	**83.01**
Never married and household income in the 1st quartile	0.30 ([0.18, 0.43])	1.35 ([1.19, 1.54])	1.35 ([1.19, 1.54])	1438	25.69
Never married and average level of education	0.25 ([0.12, 0.37])	1.28 ([1.13, 1.45])	1.28 ([1.13, 1.45])	871	13.59
Never married and personal income in the 2nd quartile	0.27 ([0.15, 0.4])	1.31 ([1.16, 1.49])	1.04 ([0.93, 1.17])	259	20.07
Unfit for work and personal income in the 2nd quartile	**−0.38 ([−0.53, −0.23])**	**0.68 ([0.59, 0.8])**	**1.98 ([1.65, 2.38])**	**234**	**35.58**
Education unknown and physical healthcare costs between €1 and €5000	**0.28 ([0.16, 0.41])**	**1.32 ([1.17,1.51])**	**1.21 ([0.95, 1.54])**	**833**	**35.53**

person years ending in suicide (N = 1,691) and one percent of all other person years (N = 1,375,966) and is used to evaluate the performance of the final model.

Step 2: the algorithm identifies significant interactions between features of interest in the training dataset. For details, see Appendix A. In short, the algorithm defines a main-effects logistic regression model including all features listed in Table 1 (hereafter referred to as basic features). Next, interaction terms are added in an iterative manner. The algorithm looks at combinations of the form "*X* and *Y*", where *X* is a feature already present in the model, and *Y* is a basic feature. So the new combination feature "*X* and *Y*" would have value 1 if both feature *X* and feature *Y* have value 1. For each of these combinations, it calculates the rate at which it would improve the log-likelihood. Then we corrected for sub-population size, since larger sub-populations without an underlying effect on suicide risk will still have a large effect on log-likelihood simply due to variance. The significant interactions that came out of this analysis were listed and for the further analyses we focused on interactions of features that had the largest effects and also included at least 200 suicides. This was done because for suicide prevention interventions the primary interest is in sub-populations with a substantial number of suicides. After this, a check was performed to ascertain whether this (interaction of) feature(s) truly improved the model. If it did not, it was removed. The process was stopped when the ratio at which removals needed to be performed exceeded 10% and at least 30 interactions were tested.

Step 3: a logistic regression model was estimated on the validation dataset including all significant interactions detected in step two. As the data in the validation set is disjoint from the training set, the notion of over-fitting is removed and regular test statistics such as t-tests and p-values can be interpreted.

Step 4: the following performance statistics were computed on the test set: log-likelihood as an indicator of model fit, and area under the receiver operating characteristics curve (AUC) as an indicator of the model's ability to distinguish between those who died by suicide and those who did not.

2.4. Statistics

For each significant feature of interest and interaction between two or more features of interest, we report the logistic regression model β parameters, odds ratios and corresponding confidence intervals. For interaction terms, we also report the compound odds ratios (CORs) and their confidence intervals, reflecting the summed effect of features when combined (e.g., $\exp(\beta_{male} + \beta_{widowed} + \beta_{male\ and\ widowed})$). Also reported are the number of suicides in the corresponding sub-populations for the validation set as well as the relative rate in said sets (per 100,000 inhabitants per year), which are corrected for the sampling procedure (number of suicides is scaled up by a factor of 2.5, and number of non-suicides by a factor of 100).

3. Results

3.1. Main effects

For a complete list of main effects predicting death by suicide, see Appendix B. Most important risk factors for suicide were middle age ($\beta_{40-54\ vs\ 25-39}$ = 0.48, 95% CI = [0.39, 0.57], OR = 1.62, 95% CI = [1.48, 1.77]; $\beta_{55-69\ vs\ 25-39}$ = 0.37, 95% CI = [0.22, 0.52], OR = 1.45, 95% CI = [1.25, 1.68]), living alone ($\beta_{living\ alone\ vs\ couple\ without\ children}$ = 0.88, 95% CI = [0.77, 0.98], OR = 2.41, 95% CI = [2.16, 2.51]), high healthcare costs ($\beta_{5-10k/year\ vs\ none}$ = 0.87, 95% CI = [0.64, 1.11], OR = 2.39, 95% CI =[1.90, 3.03]; $\beta_{>10k/year\ vs\ none}$ = 1.53, 95% CI = [1.26, 1.80], OR = 4.62, 95% CI [3.53, 6.05]), being divorced ($\beta_{divorced\ vs\ never\ married}$ = 0.51, 95% CI = [0.39, 0.62], OR = 1.67, 95% CI [1.48, 1.86]), and receiving benefits ($\beta_{short-term\ unemployment\ vs\ not}$ = 0.19, 95% CI [0.08, 0.30], OR = 1.21, 95% CI = [1.08, 1.35]; $\beta_{long-term\ unemployment\ vs\ not}$ = 0.54, 95% CI = [0.42, 0.67], OR = 1.72, 95% CI = [1.52, 1.95]; $\beta_{unfit\ for\ work\ vs\ not}$ = 1.30, 95% CI [1.16, 1.44], OR = 3.67, 95% CI = [3.19, 4.22]). Most important protective factors for suicide were female sex ($\beta_{female\ vs\ male}$ = −0.83, 95% CI = [−0.90, −0.76], OR = 0.44, 95% CI = [0.41, 0.47]), younger age ($\beta_{10-24\ vs\ 25-39}$ = −0.85, 95% CI = [−1.00, −0.71], OR = 0.43, 95% CI = [0.37, 0.49]), non-western migration background ($\beta_{first\ generation\ non-western\ vs\ Dutch}$ = −1.02, 95% CI = [−1.15, −0.89], OR = 0.36, 95% CI = [0.32, 0.41]; $\beta_{second\ generation\ non-western\ vs\ Dutch}$ = −0.53, 95% CI = [−0.70, −0.35], OR = 0.59, 95% CI = [0,50, 0.70]) and higher income (e.g. $\beta_{personal\ income\ in\ 4th\ quartile\ vs\ 1st\ quartile}$ = −0.62, 95% CI = [−0.73, −0.50], OR = 0.54, 95% CI = [0.48,0.61]). For confidence intervals of the differences between non-reference groups (i.e. 40–54 vs 10–24), see Appendix C. Among the general population there is a suicide rate of 11.8 per 100,000. When considering relative suicide rates among the sub-populations corresponding to the various features, the highest rate among the basic features is among the people who are unfit for work with a suicide rate of 47.0 per 100,000 on the validation set, with the second highest rate being among the long-term unemployed with a suicide rate of 32.1 per 100,000 on the validation set, and the rest of the sub-populations having rates below 30.0 per 100,000.

3.2. Interaction effects

Table 2 lists all twenty interaction terms included in the final logistic regression model. Of those, seventeen yielded significant effects in the validation dataset ($p < 0.05$). Among the interaction features there are ten sub-populations identified with relative risks higher than 30.0 per 100,000 on the validation set.

Broadly, three categories of interacting risk factors can be distinguished (with minor crossover): (1) interactions related to age, (2) interactions related to sex, and (3) interactions related to marital status. Two significant interactions did not fit any of these categories.

Interactions involving age: among people of young working age (25–39 years old), but not in the other age groups, a low level of education is an important risk factor for suicide (OR = 1.58 (95% CI OR [1.35,1.86], COR = 1.63 [1.38,1.93])). In contrast, being unemployed is an important risk factor for suicide in the general population but not among people of middle age (40–54 years old; OR = 0.80 (95% CI OR [0.67,0.96], COR = 2.23 [1.90,2.61])). Among those aged between 55–69, having never been married is an important risk factor (OR = 1.38 (95% CI OR [1.16,1.65], COR = 2.27 [1.64,2.44])), while high healthcare costs (OR = 0.64 (95% CI OR [0.53,0.78], COR = 4.30 [3.16,5.86])) and living alone (OR = 0.66 (95% CI OR [0.51,0.84], COR = 2.27 [1.78,2.9])) are less of a risk factor in this age group compared to other age groups (though they do remain risk factors). High healthcare costs are also less important for persons aged 70 or older (OR = 0.52 (95% CI OR [0.41,0.64], COR = 2.14 [1.58,2.90])).

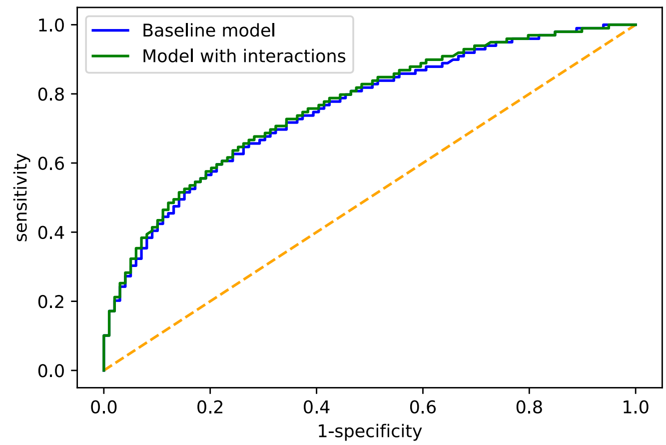


Fig. 1. Receiver Operating Characteristics curve for the baseline and the interaction models, sensitivity is the true positive rate while 1-specificity is the false positive rate. The plot shows their values for a range of thresholds.

Interactions involving sex: although being widowed is not a risk factor in general (OR = 0.91 (95% CI OR [0.76,1.10])) it is a major one for males (OR = 1.72 (95% CI OR [1.4,2.09], COR = 1.56 [1.31,1.86])). Being a part of a couple with a child at home is very protective in general (OR = 0.43 (95% CI OR [0.37,0.51])), however this effect is greatly reduced for males (OR = 1.90 (95% CI OR [1.61,2.22], COR = 0.82 [0.73,0.92])) although it does remain a protective factor.

Being on unfit for work benefits is a larger risk factor for females (OR = 3.67 (95% CI OR [3.18,4.23])) than it is for males (OR = 0.68 (95% CI OR [0.59,0.78], COR = 2.48 [2.21,2.79])). Having higher healthcare costs (€10001 or more) is a larger risk factor for females (OR = 4.62 (95% CI OR [3.54,6.05])) than it is for males (OR = 0.74 (95% CI OR [0.63,0.87], COR = 3.42 [2.64,4.43])).

Interactions involving marital status: although never being married is protective in general, in specific groups it is a risk factor: those unfit for work with low healthcare costs (OR = 1.72 (95% CI OR [1.36,2.18], COR = 6.45 [4.83,8.61])), those with the 25% lowest household incomes (OR = 1.35 (95% CI OR [1.19,1.54], COR = 1.35 [1.19,1.54])), and those with an average level of education (OR = 1.28 (95% CI OR [1.13,1.45], COR = 1.28 [1.13,1.45])).

Other interactions: finally, there are two interaction features that fit into none of the three major groups. Personal income being in the 2nd quartile is most protective for those who are unfit for work, though not so protective as to completely mitigate the risk associated with being unfit for work (OR = 0.68 (95% CI OR [0.59,0.8], COR = 1.98 [1.65,2.38])). Lastly though education being unknown is a protective factor in general (OR = 0.86 (95% CI OR [0.75,0.98])) this protective effect disappears for those with low healthcare costs (OR = 1.32 (95% CI OR [1.17,1.51], COR = 1.21 [0.95,1.54])).

3.3. Model performance

The baseline logistic regression model without interaction terms had a log-likelihood of −12184.54 and an AUC of 0.75. In comparison the logistic regression model with interaction terms had a log-likelihood of −12119.24 and an AUC of 0.76. See Fig. 1 for the curves themselves.

4. Discussion

Effective suicide prevention programs include, among others, interventions targeting subgroups of people at particularly high-risk of suicide. Here, we designed a heuristic model to detect such subgroups based on interactions between risk factors, and applied it to data covering the entire population of the Netherlands. We identified three sub-populations at ultra-high risk for suicide, with relative suicide rates of 50/100,000 person years or higher. In addition, we identified several factors that when combined increase the risk of suicide, while in isolation they do not increase the risk of suicide. These risk factors would not be detected using traditional prediction models.

We identified three sub-populations at ultra-high risk of suicide, with social isolation and socio-economic hardship as common denominators. Compared to suicide rates in the general population of the Netherlands (11.8 suicides per 100,000 person years), people who were never married and unfit for work - and among them those with low healthcare costs - were up to 7.4 times more likely to die by suicide (88 suicides per 100,000 person years). Despite the relatively small size of this group in the Dutch population, in 2012–2020 more than 100 suicides (7% of all suicides within that period) occurred in this group each year. The second ultra-high risk group concerns males who are unfit for work, with 59 suicides per 100,000 person years. These findings urge professionals in regular contact with individuals receiving unfit for work benefits, including occupational healthcare professionals, community service providers and municipal workers, to pay particular attention to males and people who were never married. The third ultra-high risk group comprises individuals aged 55–69, who were never married, are living alone and have a relatively low income, with 57 suicides per 100,000 person years. Further studies, including longitudinal and qualitative studies, are needed to investigate how the combination of these specific risk factors culminates in extreme high-risk profiles.

In addition to the extreme high-risk group, we identified several risk factors that increase the risk of suicide only in the presence of other risk factors. First, while neither young age (25–39 years old) nor lower level of education was found to be a risk factor in itself, together they constituted a major risk profile. Among individuals of young adult age, those with a lower level of education presented with a relative suicide rate more than double that of their peers with a medium or higher level of education (20.1 vs. 8.8 suicides per 100,000 person years). Our data does not provide insights into mechanisms that might underlie the elevated risk of suicide among young adults with lower education. In keeping with our prior observation that socioeconomic hardship may be a common denominator, we speculate that, among many factors, job insecurity might play a role: young adults in the Netherlands, and especially those with lower levels of education, are more likely than other age groups to be offered temporary employment [12]. Job insecurity has been linked to poorer mental health [13], which in turn is linked to a higher suicide risk [4]. To substantiate this hypothesis or find alternative explanations, we recommend research into risk factors for suicide in this group, including socio-economic factors, external stressors, psycho-social circumstances and psychological vulnerabilities.

Second, widowhood did not increase the risk of suicide in the general population in our study, yet it did when combined with the known risk factor male sex. Among widowed males, the suicide rate is more than twice the rate observed in general male population. Previous studies including males only have reported a higher risk of suicide among widowed individuals [14–16], but to our knowledge the combined risk of widowhood and male gender has not previously been reported. The current study does not allow characterisation of the suicidal process within male widowed individuals. A recent study showed that male widows, compared to female widows, are generally protected from income loss yet are more likely to experience negative emotional consequences such as loneliness and depression [17]. Our findings underline the need for social support for males who lost their partner, and urge training of gatekeepers among professionals encountering these males.

Finally, we wish to draw the readers attention to two risk factors that each appear in a large number of significant interaction terms: (1) being of middle age (55–69 years old) and (2) having never been married. The large number of significant interactions involving these factors suggests risk profiles within the sub-populations of middle-aged individuals and individuals who were never married that differ from risk profiles in the general population.

Several limitations to our approach should be considered when interpreting our findings. First, death by suicide is a relatively rare event, limiting our statistical power to find associations with risk factors. To achieve reliable model performance, we included all suicides that occurred in the Netherlands between 2012 and 2020. We are unable to assess whether results are stable over time. Second, the model is constructed bottom-up. A top-down approach starting with all possible highest-level interactions might allow detection of more high-risk subgroups, however such approaches are also known to generate more false-positives. Third, adding interaction terms to the model improved model performance only slightly (AUC = 0.76 vs. AUC = 0.75). While the validity of the identification of high-risk groups is not affected (AUC between 0.7 and 0.8 is generally deemed 'acceptable'), it does suggest that even with highly complex statistical modelling predicting death by suicide remains challenging. Fourth, we did not have data regarding family history of suicide, nor mental disorder diagnoses. These are both substantial risk factors which might explain some of the associations. Lastly, since suicide rates differ substantially across nations, there might be a limit to generalisability, especially with regard countries with substantially different cultures.

Our approach has many strengths. First, since we sampled from the entire population in a controlled manner, we avoid sampling bias. Second, our model is hypothesis-free, allowing identification of previously unidentified risk groups. Third, our model has flexible settings, allowing the user to adjust the trade-off between good model performance and statistically robust results. Finally, and in contrast to existing machine learning methods such as artificial neural networks, our model is open and readily interpretable.

4.1. Conclusions

In summary, we performed a heuristic machine learning method to find interactions. We found disproportionately high suicide rates among people who were never married and received unfit for work benefits, among males who received unfit for work benefits, and among those aged 55–69 who lives alone, were never married and whose household income was low. Additionally, we found high suicide rates among those aged 25–39 with a low level of education and among males who lost their partner. Our findings may have important implications for suicide prevention policies and are generalizable to other (similar) countries.

Funding

This paper was funded by 113 Suicide Prevention, which is in turn funded by the Dutch Ministry of Health, Welfare, and Sport.

Declaration of Competing Interest

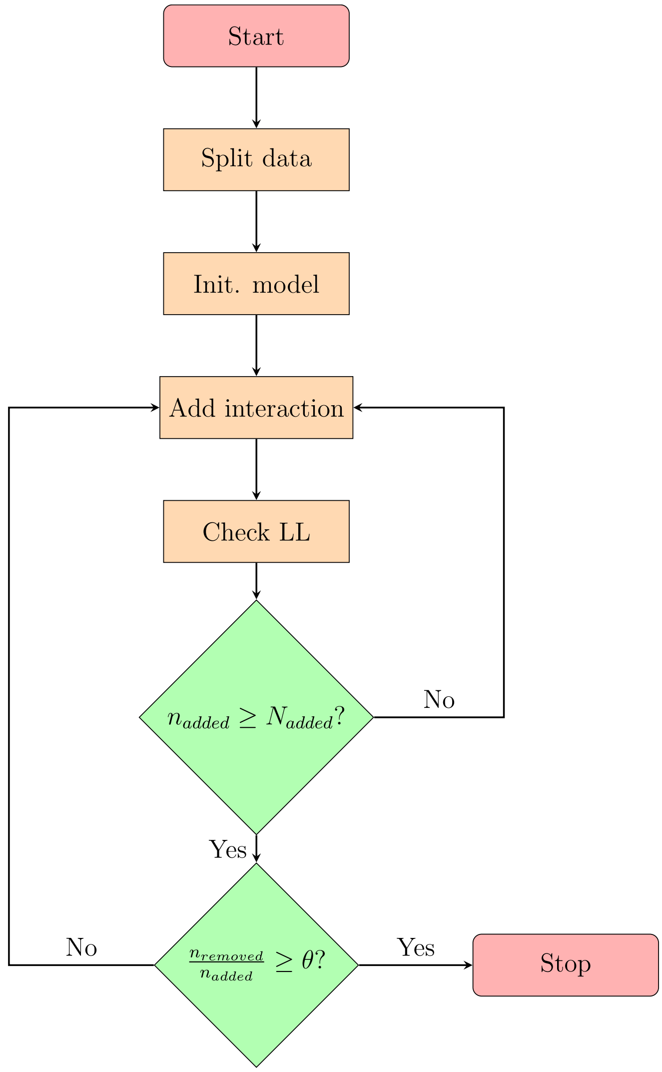
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

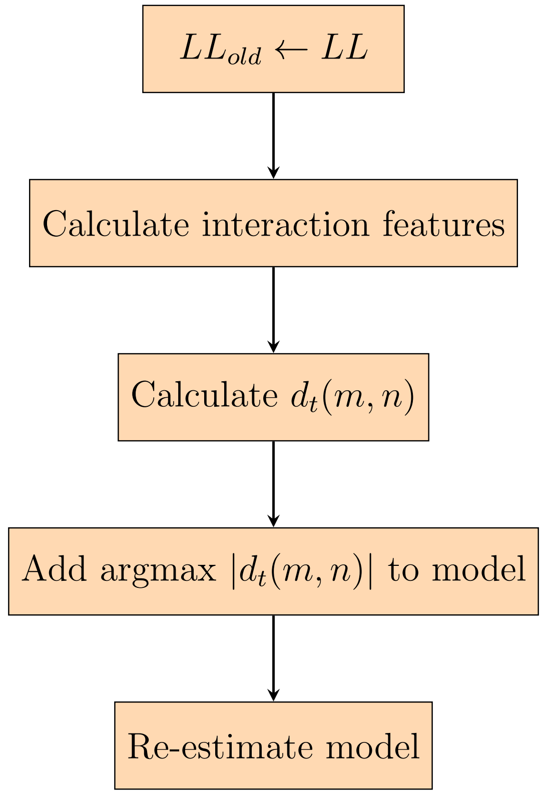
Results based on calculations by 113 Suicide Prevention using non-public microdata from Statistics Netherlands. Under certain conditions, these microdata are accessible for statistical and scientific research. For further information: microdata@cbs.nl.

Appendix A. Full explanation Step 2 algorithm

A.1. Global flowchart



A.2. Flowchart “Add Interaction” process



A.3. Elaboration flowchart

In what follows we will outline the full details of every step within the global flowchart, further splitting the “Add Interaction” step into the sub-steps shown in the second flowchart.

Start:

To start with we specify our hyper-parameters N_{added}, θ, t, and S_{min} whose functions shall be explained as they become relevant. Additionally, we initialize $n_{added} = n_{removed} = 0$ and T as an empty list. These will be updated throughout the procedure.

We define $\overrightarrow{x}_i$ for $i \in \{1, 2, \dots, N\}$ to be our one-hot encoded basic features. We define $\overrightarrow{y}_i$ for $i \in \{1, 2, \dots, L\}$ to be all the features in our model. The amount of basic features, N, is fixed. However, since we will be adding features throughout our model, the total amount of features, L, will vary.

Split data:

We split our training set into two subsets: a *searching* set (80% of cases), and a *control* set (containing the remaining 20%).

Init. model:

Using the searching set we estimate an initial logistic regression model specified by

$$\mathbb{P}((\overrightarrow{s})_k = 1 | \overrightarrow{y_1}, \dots, \overrightarrow{y_L}) = \frac{e^{V_k}}{1 + e^{V_k}}$$

where $\overrightarrow{s}$ is the feature corresponding to “died by suicide” and

$$V_k(\overrightarrow{y}_1, \dots, \overrightarrow{y}_L) = \beta_0 + \sum_{i=1}^{L} \beta_i (\overrightarrow{y}_i)_k$$

with the β_i being the parameters to be estimated. Estimation is done through log-likelihood maximization via gradient descent methods. Set LL to be equal to the log-likelihood of the model on the control set.

Add interaction:

$LL_{old} \leftarrow LL$: We set the value of LL_{old} to the current value of LL.

Calculate interaction features: For each $m \in \{1, \dots, N\}$ and $n \in \{1, \dots, L\}$ define $\overrightarrow{z}_{m,n} = \overrightarrow{x}_m * \overrightarrow{y}_n$ where * denotes the element-wise product.

Let $\overrightarrow{u}$ be the all ones vector and $N_{\overrightarrow{z}_{m,n}} = \langle \overrightarrow{z}_{m,n}, \overrightarrow{u} \rangle$ be the amount of people possessing both characteristic *m and n*. Let $S_{\overrightarrow{z}_{m,n}} = \langle \overrightarrow{z}_{m,n}, \overrightarrow{s} \rangle$ be the amount of people possessing both characteristic *m and n* who died by suicide.

Let $s_{\overrightarrow{z}_{m,n}} = \mathbb{1}(\mathbb{S}_{\overrightarrow{\mathbb{Z}}_{m,n}} \geqslant \mathbb{S}_{\min})$. Here S_{min} functions as a lower bound on the amount of suicides in the sub-population corresponding to the interaction feature for us to consider it for the model. We used $S_{min} = 200$.

Calculate $d_t(m,n)$: Let $LL_{m,n}(\beta_{m,n})$ be the log-likelihood corresponding to the logistic regression model specified as

Table B.3
Full results logistic regression on validation set including both basic features and interaction terms. With corresponding Beta parameters, Odds-Ratios, Compound Odds Ratios, absolute and relative number of suicides within the sub-population within the validation set as well as the training set. With N(val)=absolute number of suicides within validation set, N(train)=absolute number of suicides within training set, Rel(val) = relative number of suicides within the validation set (corrected for sampling procedure, per 100,000), Rel(train)=relative number of suicides within the training set (corrected for sampling procedure, per 100,000).

Features	Beta estimates	95% C.I. Beta	95% C.I. OR	95% C.I. COR	N (val)	Rel (val)	N (train)	Rel (train)
β_0/ Full population	−5.42	[−5.7,−5.13]	[0,0.01]	[0,0.01]	6512	11.7598	8214	11.8591
Male	0.00	[0,0]	[1,1]	[1,1]	4397	16.0445	5565	16.2555
Aged 25–39	0.00	[0,0]	[1,1]	[1,1]	1151	10.0758	1467	10.2593
Dutch Immigration Background	0.00	[0,0]	[1,1]	[1,1]	5378	12.4660	6756	12.5191
Part couple without child at home	0.00	[0,0]	[1,1]	[1,1]	1510	9.4118	1857	9.2573
Personal income in first quartile	0.00	[0,0]	[1,1]	[1,1]	941	6.9806	1228	7.3130
Household income in first quartile	0.00	[0,0]	[1,1]	[1,1]	2784	20.3763	3401	19.9394
Household wealth/debts in first quartile	0.00	[0,0]	[1,1]	[1,1]	1814	13.3128	2176	12.8107
Average level of education	0.00	[0,0]	[1,1]	[1,1]	1448	12.5832	1896	13.2622
No physical healthcare costs	0.00	[0,0]	[1,1]	[1,1]	86	10.8723	123	12.2103
Never married	0.00	[0,0]	[1,1]	[1,1]	2821	12.3053	3508	12.2679
Female	−0.83	[−0.9,−0.76]	[0.41,0.47]	[0.41,0.47]	2115	7.5616	2649	7.5623
Aged 10–24	−0.85	[−1,−0.71]	[0.37,0.49]	[0.37,0.49]	512	4.5826	720	5.1680
Aged 40–54	0.48	[0.39,0.57]	[1.48,1.76]	[1.48,1.76]	1956	15.7218	2403	15.4614
Aged 55–69	0.37	[0.22,0.52]	[1.24,1.68]	[1.24,1.68]	1796	15.3007	2231	15.1825
Aged 70 or older	−0.11	[−0.24,0.03]	[0.79,1.03]	[0.79,1.03]	928	12.0496	1202	12.4417
1st generation western immigration background	−0.21	[−0.33,−0.09]	[0.72,0.92]	[0.72,0.92]	331	11.6500	396	11.0959
1st generation non-western immigration background	−1.02	[−1.15,−0.89]	[0.32,0.41]	[0.32,0.41]	297	7.0322	359	6.8358
2nd generation western immigration background	−0.06	[−0.17,0.06]	[0.84,1.06]	[0.84,1.06]	363	13.0852	493	14.2122
2nd generation non-western immigration background	−0.53	[−0.7,−0.35]	[0.5,0.7]	[0.5,0.7]	143	5.9703	210	6.9805
Child living at home	0.08	[−0.08,0.24]	[0.93,1.27]	[0.93,1.27]	508	4.9926	756	5.9679
Living alone	0.88	[0.77,0.98]	[2.17,2.66]	[2.17,2.66]	2943	27.4229	3652	27.2016
Part couple with child at home	−0.84	[−1,−0.68]	[0.37,0.51]	[0.37,0.51]	1052	7.1662	1341	7.2863
Other member household	0.14	[0.01,0.27]	[1.01,1.32]	[1.01,1.32]	499	13.3264	608	12.9204
Personal income in the 2nd quartile	−0.23	[−0.35,−0.12]	[0.71,0.89]	[0.71,0.89]	2184	15.9142	2734	15.9101
Personal income in the 3rd quartile	−0.42	[−0.52,−0.32]	[0.6,0.73]	[0.6,0.73]	1847	13.6120	2305	13.5711
Personal income in the 4th quartile	−0.62	[−0.73,−0.5]	[0.48,0.61]	[0.48,0.61]	1407	10.3917	1782	10.5031
Personal income unknown	0.20	[−0.03,0.42]	[0.97,1.53]	[0.97,1.53]	133	12.5132	165	12.3466
Household income in the 2nd quartile	0.00	[−0.1,0.09]	[0.91,1.1]	[0.91,1.1]	1588	11.6848	2057	12.1188
Household income in the 3rd quartile	−0.04	[−0.16,0.07]	[0.86,1.07]	[0.86,1.07]	1142	8.4459	1384	8.1440
Household income in the 4th quartile	−0.20	[−0.32,−0.07]	[0.72,0.94]	[0.72,0.94]	865	6.3886	1207	7.1401
Household net wealth in the 2nd quartile	−0.05	[−0.12,0.02]	[0.89,1.02]	[0.89,1.02]	1848	13.5832	2387	14.0547
Household net wealth in the 3rd quartile	−0.02	[−0.1,0.06]	[0.9,1.06]	[0.9,1.06]	1336	9.8571	1657	9.7626
Household net wealth in the 4th quartile	0.10	[0.02,0.19]	[1.02,1.21]	[1.02,1.21]	1381	10.2071	1829	10.7673
Low level of education	0.03	[−0.09,0.14]	[0.92,1.15]	[0.92,1.15]	1248	10.4582	1478	9.9127
High level of education	0.03	[−0.08,0.14]	[0.92,1.16]	[0.92,1.16]	893	9.8205	1065	9.2930
Level of education unknown	−0.15	[−0.29,−0.02]	[0.75,0.98]	[0.75,0.98]	2923	12.7969	3775	13.2008
Physical healthcare costs between €1 and €5000	0.06	[−0.17,0.28]	[0.84,1.33]	[0.84,1.33]	5053	10.4067	6374	10.5056
Physical healthcare costs between €5001 and €10000	0.87	[0.64,1.11]	[1.89,3.02]	[1.89,3.02]	587	21.8782	727	21.5478
Physical healthcare costs of €10001 or more	1.53	[1.26,1.8]	[3.54,6.05]	[3.54,6.05]	786	23.4980	990	23.5258
Physical healthcare costs unknown	−1.40	[−1.69,−1.11]	[0.18,0.33]	[0.18,0.33]	71	7.9273	78	6.9189
Married or registered partnership	0.26	[0.14,0.37]	[1.15,1.45]	[1.15,1.45]	2096	8.5726	2608	8.5051
Divorced	0.51	[0.39,0.62]	[1.48,1.86]	[1.48,1.86]	1155	24.6854	1489	25.5291
Widowed	−0.09	[−0.27,0.09]	[0.76,1.1]	[0.76,1.1]	440	14.2491	609	15.6787
Short-term unemployment	0.19	[0.08,0.3]	[1.09,1.35]	[1.09,1.35]	395	15.8520	503	16.1755
Unfit for work	1.30	[1.16,1.44]	[3.18,4.23]	[3.18,4.23]	1048	46.9534	1262	44.8177
Long-term unemployment	0.54	[0.42,0.67]	[1.52,1.95]	[1.52,1.95]	609	32.0567	746	31.2095
Aged 25–39 and low level of education	0.46	[0.3,0.62]	[1.35,1.86]	[1.38,1.93]	259	20.0663	296	18.2429
Aged 40–54 and long-term unemployment	−0.22	[−0.41,−0.04]	[0.67,0.96]	[1.9,2.61]	234	35.5796	262	31.7326
Aged 55–69 and living alone	−0.42	[−0.67,−0.17]	[0.51,0.84]	[1.78,2.9]	833	35.5369	1040	35.6329
Aged 55–69 and living alone and Dutch immigration background	0.18	[−0.04,0.39]	[0.96,1.48]	[2.3,3.19]	728	39.3718	892	38.8586
Aged 55–69 and living alone and household income in the 1st quartile and never married	−0.21	[−0.43,0.01]	[0.65,1.01]	[2.6,4.55]	229	57.2214	250	50.4134
Aged 55–69 and never married	0.32	[0.15,0.5]	[1.16,1.65]	[1.64,2.44]	427	34.8185	506	33.1695
Aged 55–69 and part of couple without child at home	−0.46	[−0.63,−0.29]	[0.53,0.75]	[0.79,1.05]	622	9.3768	753	9.0842
Aged 55–69 and healthcare costs of €10001 or more	−0.44	[−0.63,−0.25]	[0.53,0.78]	[3.16,5.86]	238	30.7018	280	29.0080
Aged 70 or older and healthcare costs of €10001 or more	−0.66	[−0.88,−0.44]	[0.41,0.64]	[1.58,2.9]	175	15.5938	260	18.4981
Male and unfit for work	−0.39	[−0.54,−0.24]	[0.59,0.78]	[2.21,2.79]	642	58.5574	764	55.5414
Male and part of couple with child at home	0.64	[0.48,0.8]	[1.61,2.22]	[0.73,0.92]	801	10.9391	979	10.6842
Male and widowed	0.54	[0.33,0.74]	[1.4,2.09]	[1.31,1.86]	218	31.3128	304	34.5278
Male and healthcare costs of €10001 or more	−0.30	[−0.46,−0.14]	[0.63,0.87]	[2.64,4.43]	456	27.4831	596	28.4100
Never married and unfit for work	−0.03	[−0.26,0.19]	[0.77,1.21]	[2.77,4.53]	441	88.4831	495	79.0293
Never married and unfit for work and physical healthcare costs between €1 and €5000	0.54	[0.31,0.78]	[1.36,2.18]	[4.83,8.61]	321	83.0144	362	74.6546
Never married and household income in the 1st quartile	0.30	[0.18,0.43]	[1.19,1.54]	[1.19,1.54]	1438	25.6896	1715	24.6509
Never married and average level of education	0.25	[0.12,0.37]	[1.13,1.45]	[1.13,1.45]	871	13.5912	1144	14.3792
Never married and personal income in the 2nd quartile	0.27	[0.15,0.4]	[1.16,1.49]	[0.93,1.17]	1008	24.7583	1245	24.5072
Unfit for work and personal income in the 2nd quartile	−0.38	[−0.53,−0.23]	[0.59,0.8]	[1.65,2.38]	382	48.5758	470	47.5203
Education unknown and physical healthcare costs between €1 and €5000	0.28	[0.16,0.41]	[1.17,1.51]	[0.95,1.54]	2165	11.5392	2808	11.9722

Table C.4
Differences of beta parameters of the age groups with corresponding 95% confidence intervals (significant differences are marked with a *).

AB	Age 10–24	Age 25–39	Age 40–54	Age 55–69	Age 70+
Age 10–24	N/A	−0.85 [−1.00,−0.71]*	−1.33 [−1.48,−1.18]*	−1.22 [−1.41,−1.03]*	−0.74 [−0.92,−0.56]*
Age 25–39	0.85 [0.71,1.00]*	N/A	−0.48 [−0.57,−0.39]*	−0.37 [−0.52,−0.22]*	0.11 [−0.03,0.24]
Age 40–54	1.33 [1.18,1.48]*	0.48 [0.39,0.57]*	N/A	0.11 [−0.02,0.24]	0.59 [0.47,0.71]*
Age 55–69	1.22 [1.03,1.41]*	0.37 [0.22,0.52]*	−0.11 [−0.24,0.02]	N/A	0.48 [0.32,0.64]*
Age 70+	0.74 [0.56,0.92]*	−0.11 [−0.24,0.03]	−0.59 [−0.71,−0.47]*	0.48 [0.32,0.64]*	N/A

Table C.5
Differences of beta parameters of the migration backgrounds with corresponding 95% confidence intervals (significant differences are marked with a *).

AB	Dutch	1st gen Western	1st gen non-Western	2nd gen Western	2nd gen non-Western
Dutch	N/A	0.21 [0.09,0.33]*	1.02 [0.89,1.15]*	0.06 [−0.06,0.17]	0.53 [0.35,0.7]*
1st gen Western	−0.21 [−0.33,−0.09]*	N/A	0.81 [0.65,0.97]*	−0.15 [−0.31,0.01]	0.32 [0.12,0.52]*
1st gen non-Western	−1.02 [−1.15,−0.89]*	−0.81 [−0.97,−0.65]*	N/A	−0.96 [−1.12,−0.80]*	−0.49 [−0.69,−0.29]*
2nd gen Western	−0.06 [−0.17,0.06]	0.15 [−0.01,0.31]	0.96 [0.80,1.12]*	N/A	0.47 [0.27,0.67]*
2nd gen non-Western	−0.53 [−0.7,−0.35]*	−0.32 [−0.52,−0.12]*	0.49 [0.29,0.69]*	−0.47 [−0.67,−0.27]*	N/A

Table C.6
Differences of beta parameters of place in household with corresponding 95% confidence intervals (significant differences are marked with a *).

AB	Child living at home	Living alone	Partner couple without kids	Partner couple with kids	Other
Child living at home	N/A	−0.80 [−0.94,−0.66]*	0.08 [−0.08,0.24]	0.92 [0.72,1.12]*	−0.06 [−0.22,0.10]
Living alone	0.80 [0.66,0.94]*	N/A	0.88 [0.77,0.98]*	1.72 [1.56,1.88]*	0.74 [0.63,0.85]*
Partner couple without kids	−0.08 [−0.24,0.08]	−0.88 [−0.98,−0.77]*	N/A	0.84 [0.68,1.00]*	−0.14 [−0.27,−0.01]*
Partner couple with kids	−0.92 [−1.12,−0.72]*	−1.72 [−1.88,−1.56]*	−0.84 [−1,−0.68]*	N/A	−0.98 [−1.15,−0.81]*
Other	0.06 [−0.10,0.22]	−0.74 [−0.85,−0.63]*	0.14 [0.01,0.27]*	0.98 [0.81,1.15]*	N/A

$$V_k = \beta_0 + \sum_{i=1}^{L} \beta_i (\overrightarrow{y}_i)_k + \beta_{m,n} (\overrightarrow{z}_{m,n})_k$$

then

$$\frac{dLL_{m,n}}{\beta_{m,n}} = \sum_{k=1}^{N_p} (\overrightarrow{z}_{m,n})_k \left(s_k - \frac{e^{V_k}}{1+e^{V_k}} \right)$$

where N_p is the total number of cases in our searching set. Note that under the assumption that the "true" value of $\beta_{n,m}$ on the underlying probability process is 0 (i.e. feature $\overrightarrow{z}_{m,n}$ is irrelevant) the value of this expression scales to the order of $\sqrt{N_{\overrightarrow{z}_{m,n}}}$. Therefore, if we do not correct for this, large values of $|\frac{dLL_{m,n}}{\beta_{m,n}}|$ will simply end up corresponding to large sub-populations. As such we define

$$d_t(m,n) = \frac{1}{N^t_{\overrightarrow{z}_{m,n}}} \frac{dLL_{m,n}}{\beta_{m,n}} s_{\overrightarrow{z}_{m,n}}$$

where hyper-parameter t describes the trade-off between optimization of the log-likelihood and statistical significance, with a value of 0 completely prioritizing the former, and a value of 0.5 completely prioritizing the latter. We used $t = 0.3$.

Add $\text{argmax}|d_t(m,n)|$ *to model:* We then select

$$(m^*, n^*) = \underset{m,n}{\text{argmax}} |d_t(m,n)|$$

and add the corresponding feature to our model by setting $\overrightarrow{y}_{L+1} = \overrightarrow{z}_{m^*,n^*}$ and set $L \leftarrow L+1$. We add (m^*, n^*) to the list T. We also set $n_{added} \leftarrow n_{added} + 1$.

Re-estimate model: We re-estimate the model with the new feature and set LL to the log-likelihood of this new model on the control set.

Check LL:

We check whether or not the performance on the control set has improved by looking at $LL - LL_{old}$. If this is negative we once again remove the added feature from our model and set $n_{removed} \leftarrow n_{removed} + 1 \mathbf{n_{added}} \geqslant N_{added}$:

Here N_{added} functions as a minimum number of iterations before stopping. If we have not yet run that many iterations, we return to the "Add interaction" step. If we have we move on to the next step. We used $N_{added} = 30$.

$\frac{\mathbf{n_{removed}}}{\mathbf{n_{added}}} \geqslant \theta$:

Here θ functions as a minimum amount of false positives before terminating. If the proportion of false positives is less that θ we return to the "Add interaction" step. If it is at least θ we end our algorithm. We used $\theta = 0.1$.

Table C.7
Differences of beta parameters of personal income with corresponding 95% confidence intervals (significant differences are marked with a *).

AB	1st quartile	2nd quartile	3rd quartile	4th quartile	Unknown
1st quartile	N/A	0.23 [0.12,0.35]*	0.42 [0.32,0.52]*	0.62 [0.5,0.73]*	−0.20 [−0.42,0.03]
2nd quartile	−0.23 [−0.35,−0.12]*	N/A	0.19 [0.10,0.28]*	0.39 [0.28,0.50]*	−0.43 [−0.65,−0.21]*
3rd quartile	−0.42 [−0.52,−0.32]*	−0.19 [−0.28,−0.10]*	N/A	0.20 [0.12,0.28]*	−0.62 [−0.84,−0.40]*
4th quartile	−0.62 [−0.73,−0.5]*	−0.39 [−0.50,−0.28]*	−0.20 [−0.28,−0.12]*	N/A	−0.82 [−1.05,−0.59]*
Unknown	0.20 [−0.03,0.42]	0.43 [0.21,0.65]*	0.62 [0.40,0.84]*	0.82 [0.59,1.05]*	N/A

Table C.8
Differences of beta parameters of household income with corresponding 95% confidence intervals (significant differences are marked with a *).

A\B	1st quartile	2nd quartile	3rd quartile	4th quartile
1st quartile	N/A	0.00 [−0.09,0.10]	0.04 [−0.16,0.07]	0.20 [0.07,0.32]*
2nd quartile	0.00 [−0.10,0.09]	N/A	0.04 [−0.04,0.12]	0.20 [0.10,0.30]*
3rd quartile	−0.04 [−0.16,0.07]	−0.04 [−0.12,0.04]	N/A	0.16 [0.07,0.25]*
4th quartile	−0.20 [−0.32,−0.07]*	−0.20 [−0.30,−0.10]*	−0.16 [−0.25,−0.07]*	N/A

Table C.9
Differences of beta parameters of net household wealth with corresponding 95% confidence intervals (significant differences are marked with a *).

A\B	1st quartile	2nd quartile	3rd quartile	4th quartile
1st quartile	N/A	0.05 [−0.02,0.12]	0.02 [−0.06,0.10]	−0.10 [−0.19,−0.02]*
2nd quartile	−0.05 [−0.12,0.02]	N/A	−0.03 [−0.11,0.05]	−0.15 [−0.23,−0.07]*
3rd quartile	−0.02 [−0.10,0.06]	0.03 [−0.05,0.11]	N/A	−0.12 [−0.20,−0.04]*
4th quartile	0.10 [0.02,0.19]*	0.15 [0.07,0.23]*	0.12 [0.04,0.20]*	N/A

Table C.10
Differences of beta parameters of education level with corresponding 95% confidence intervals (significant differences are marked with a *).

A\B	Low	Mid	High	Unknown
Low	N/A	0.03 [−0.09,0.14]	0.00 [−0.10,0.10]	0.18 [0.05,0.31]*
Mid	−0.03 [−0.14,0.09]	N/A	−0.03 [−0.14,0.08]	0.15 [0.02,0.29]*
High	0.00 [−0.10,0.10]	0.03 [−0.08,0.14]	N/A	0.18 [0.05,0.31]*
Unknown	−0.18 [−0.31,−0.05]*	−0.15 [−0.29,−0.02]*	−0.18 [−0.31,−0.05]*	N/A

Table C.11
Differences of beta parameters of physical healthcare costs with corresponding 95% confidence intervals (significant differences are marked with a *).

AB	€0	€1–5000	€5001–10000	€10001+	Unknown
€0	N/A	−0.06 [−0.28,0.17]	−0.87 [−1.11,−0.64]*	−1.53 [−1.80,−1.26]*	1.40 [1.11,1.69]*
€1–5000	0.06 [−0.17,0.28]	N/A	−0.81 [−0.92,−0.70]*	−1.47 [−1.63,−1.31]*	1.46 [1.07,1.85]*
€5001–10000	0.87 [0.64,1.11]*	0.81 [0.70,0.92]*	N/A	−0.66 [−0.83,−0.49]*	2.27 [1.88,2.66]*
€10001+	1.53 [1.26,1.80]*	1.47 [1.31,1.63]*	0.66 [0.49,0.83]*	N/A	2.93 [2.49,3.37]*
Unknown	−1.40 [−1.69,−1.11]*	−1.46 [−1.85,−1.07]*	−2.27 [−2.66,−1.88]*	−2.93 [−3.37,−2.49]*	N/A

Table C.12
Differences of beta parameters of marital status with corresponding 95% confidence intervals (significant differences are marked with a *).

AB	Never married	Married	Divorced	Widowed	Unknown
Never married	N/A	−0.26 [−0.37,−0.14]*	−0.51 [−0.62,−0.39]*	0.09 [−0.09,0.27]	1.40 [1.11,1.69]*
Married	0.26 [0.14,0.37]*	N/A	−0.25 [−0.35,−0.15]*	0.35 [0.18,0.52]*	1.46 [1.07,1.85]*
Divorced	0.51 [0.39,0.62]*	0.25 [0.15,0.35]*	N/A	0.60 [0.44,0.76]*	2.27 [1.88,2.66]*
Widowed	−0.09 [−0.27,0.09]	−0.35 [−0.52,−0.18]*	−0.60 [−0.76,−0.44]*	N/A	2.93 [2.49,3.37]*
Unknown	−1.40 [−1.69,−1.11]*	−1.46 [−1.85,−1.07]*	−2.27 [−2.66,−1.88]*	−2.93 [−3.37,−2.49]*	N/A

Appendix B. Full results logistic regression

In Table B.3 we give the full results of our final model including both the basic as well as the interaction features.

Appendix C. Confidence intervals differences non-reference groups ($\beta_A - \beta_B$)

It is interesting to not only know whether or not sub-populations have an increased risk of suicide with respect to a reference sub-population, but also with respect to the other sub-populations. Therefore, we provide confidence intervals for $\beta_A - \beta_B$ for sub-populations corresponding to the same original categorical variable in Tables C.4 to C.12.

References

[1] Berkelmans G, van der Mei R, Bhulai S, Gilissen R. Identifying socio-demographic risk factors for suicide using data on an individual level. BMC Public Health 2021; 21(1):1702. https://doi.org/10.1186/s12889-021-11743-3.

[2] World Health Organization. Preventing suicide: a global imperative. Geneva: World Health Organization; 2014. https://apps.who.int/iris/handle/10665/131056.

[3] Ayhan G, Arnal R, Basurko C, About V, Pastre A, Pinganaud E, Sins D, Jehel L, Falissard B, Nacher M. Suicide risk among prisoners in French Guiana: prevalence and predictive factors. BMC Psychiatry 2017;17(1):156. https://doi.org/10.1186/s12888-017-1320-4. http://bmcpsychiatry.biomedcentral.com/articles/10.1186/s12888-017-1320-4.

[4] Bhatt M, Perera S, Zielinski L, Eisen RB, Yeung S, El-Sheikh W, DeJesus J, Rangarajan S, Sholer H, Iordan E, Mackie P, Islam S, Dehghan M, Thabane L, Samaan Z. Profile of suicide attempts and risk factors among psychiatric patients: A case-control study. PLOS ONE 2018;13(2):e0192998. https://doi.org/10.1371/journal.pone.0192998. https://dx.plos.org/10.1371/journal.pone.0192998.

[5] Choi SB, Lee W, Yoon J-H, Won J-U, Kim DW. Risk factors of suicide attempt among people with suicidal ideation in South Korea: a cross-sectional study. BMC Public Health 2017;17(1):579. https://doi.org/10.1186/s12889-017-4491-5. http://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-017-4491-5.

[6] Franklin JC, Ribeiro JD, Fox KR, Bentley KH, Kleiman EM, Huang X, Musacchio KM, Jaroszewski AC, Chang BP, Nock MK. Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. Psychol Bull 2017; 143(2):187–232. https://doi.org/10.1037/bul0000084. http://doi.apa.org/getdoi.cfm?doi=10.1037/bul0000084.

[7] Parra-Uribe I, Blasco-Fontecilla H, Garcia-Parés G, Martínez-Naval L, Valero-Coppin O, Cebrià-Meca A, Oquendo MA, Palao-Vidal D. Risk of re-attempts and suicide death after a suicide attempt: A survival analysis. BMC Psychiatry 2017;17(1):163. https://doi.org/10.1186/s12888-017-1317-z. http://bmcpsychiatry.biomedcentral.com/articles/10.1186/s12888-017-1317-z.

[8] Uher R. Gene-Environment Interactions in Severe Mental Illness. Front Psychiatry May 2014;5. https://doi.org/10.3389/fpsyt.2014.00048. http://journal.frontiersin.org/article/10.3389/fpsyt.2014.00048/abstract.

[9] Gradus JL, Rosellini AJ, Horváth-Puhó E, Street AE, Galatzer-Levy I, Jiang T, Lash TL, Sørensen HT. Prediction of Sex-Specific Suicide Risk Using Machine Learning and Single-Payer Health Care Registry Data From Denmark. JAMA Psychiatry 2020;77(1):25–34. https://doi.org/10.1001/jamapsychiatry.2019.2905.

[10] Zheng L, Wang O, Hao S, Ye C, Liu M, Xia M, Sabo AN, Markovic L, Stearns F, Kanov L, Sylvester KG, Widen E, McElhinney DB, Zhang W, Liao J, Ling XB. Development of an early-warning system for high-risk patients for suicide attempt using deep learning and electronic health records. Trans Psychiatry 2020;10(1): 1–10. https://doi.org/10.1038/s41398-020-0684-2. https://www.nature.com/articles/s41398-020-0684-2.

[11] Kirtley OJ, van Mens K, Hoogendoorn M, Kapur N, de Beurs D. Translating promise into practice: a review of machine learning in suicide research and prevention. Lancet Psychiatry 2022;9(3):243–52. https://doi.org/10.1016/S2215-0366(21)00254-6. https://linkinghub.elsevier.com/retrieve/pii/S2215036621002546.

[12] Aantal flexwerkers in 15 jaar met drie kwart gegroeid (Feb. 2019). URL:https://www.cbs.nl/nl-nl/nieuws/2019/07/aantal-flexwerkers-in-15-jaar-met-drie-kwart-gegroeid.

[13] LaMontagne AD, Too LS, Punnett L, Milner AJ. Changes in Job Security and Mental Health: An Analysis of 14 Annual Waves of an Australian Working-Population Panel Survey. Am J Epidemiol 2021;190(2):207–15. https://doi.org/10.1093/aje/kwaa038. https://academic.oup.com/aje/article/190/2/207/5815383.

[14] Bower KL, Emerson KG. Exploring Contextual Factors Associated with Suicide among Older Male Farmers: Results from the CDC NVDRS Dataset. Clin Gerontol 2021;44(5):528–35. https://doi.org/10.1080/07317115.2021.1893885. https://www.tandfonline.com/doi/full/10.1080/07317115.2021.1893885.

[15] Yang J, He G, Chen S, Pan Z, Zhang J, Li Y, Lyu J. Incidence and risk factors for suicide death in male patients with genital-system cancer in the United States. Eur J Surg Oncol 2019;45(10):1969–76. https://doi.org/10.1016/j.ejso.2019.03.022. https://linkinghub.elsevier.com/retrieve/pii/S0748798319303506.

[16] Richardson C, Robb KA, O'Connor RC. A systematic review of suicidal behaviour in men: A narrative synthesis of risk factors. Soc Sci Med 2021;276:113831. https://doi.org/10.1016/j.socscimed.2021.113831. https://linkinghub.elsevier.com/retrieve/pii/S0277953621001635.

[17] Streeter JL. Gender differences in widowhood in the short run and long run: financial and emotional well-being. Innov Aging 2019;3(Suppl. 1):S736. https://doi.org/10.1093/geroni/igz038.2698. https://academic.oup.com/innovateage/article/3/Supplement_1/S736/5618394.