Supplemental material for: "A new ranking index to identify the work-related psychosocial factors most impacting mental health: a cross sectional study"

Table of Contents:

Appendix A: Description of the Weifila method

Developed by Wallard [1], the Weifila (Weighted First Last) method is based on variance decomposition used in a linear regression context. This method consists in assigning to each predictor X_i a part of variance $w(i)$ which is a weighted average between first allocation "$first(i)$" and last allocation "$last(i)$".

The "first" allocation is computed as the squared correlation of the predictor with the response: $first(i) = [r(Y; X_i)]^2$. It is the value which is considered for the selection of the first predictor in a forward algorithm.

The "last" allocation is a measure of the relative importance for a predictor i, computed as the increase in R^2 when predictor i is included last in the model compared to the R^2 with only the other $p - 1$ predictors. This measure is sometimes presented as the amount by which the R^2 is reduced when this predictor is deleted from the regression equation. To calculate the $w(i)$ terms, first the sum of $first(i)$ and $last(i)$ for all predictors are calculated: $\mathrm{F} = \sum_i first(i)$, $\mathrm{L} = \sum_i last(i)$ and two situations are considered in the usual situation where $F \neq L$:

if $L < R^2 < F$ then $w(i) = last(i)\left(\frac{F-R^2}{F-L}\right) + first(i)\left(\frac{R^2-L}{F-L}\right)$ **Equation 1**

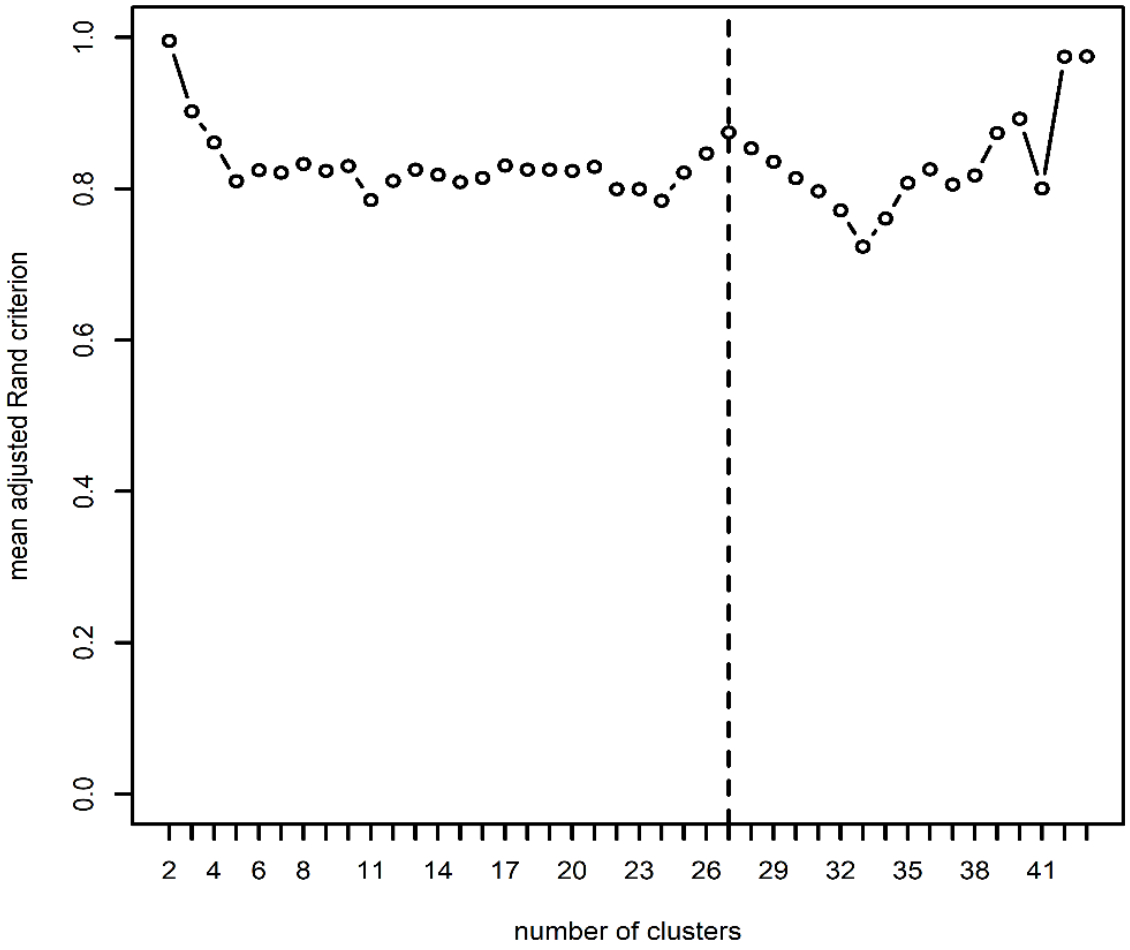
if $F < R^2 < L$ then $w(i) = last(i)\left(\frac{R^2-F}{L-F}\right) + first(i)\left(\frac{L-R^2}{L-F}\right)$ **Equation 2**

We directly used *first* and *last* options implemented in the R package **relaimpo** [2,3] to compute $w(i)$.

Appendix B: Description of the random forests method

Introduced by Breiman in 2001 [4], Random Forests are a successful machine learning approach generalizing the methodology of Classification And Regression Trees (CART) [5]. This method has been successfully applied in various applications such as genomic [6] and early detection of Alzheimer [7]. Basically, random forest trees are grown nondeterministically using several bootstrap samples from the original dataset. In this context, each tree has a set of observations from the original dataset which were not used to develop it. These are termed Out-Of-Bag (OOB) observations, and can be used to estimate the variable importance. The most frequent approach to estimate variable importance is termed permutation importance. According to this measure, a variable X_i is considered as important if it has a positive effect on the prediction performance, estimated by the OOB prediction error. To measure the importance of the variable X_i, any association between the variable and the outcome is broken by randomly permuting the values of all individuals of X_i in OOB observation. Another OOB prediction error is computed with the permuted values of X_i. The importance of variable X_i is the difference in prediction error before and after permuting X_i, averaged over all trees of the forest. The larger the permutation importance of a variable, the more relevant the variable is for the overall prediction accuracy. The R package **party** (function **cforest**) [8,9] is one of the most relevant methods implemented for computing importance of predictors in random forest. Indeed, this function gives unbiased results even if predictor variables are of different types (for example: different scale of measurement, different number of categories).

Appendix C: Cluster partition stability using bootstrap approach



Look at the peaks in the plot above. The maximum for stability is 1.0, so the higher the peak, the better. In this case, the plot seems to suggest twenty-seven stable partitions are present in the data.

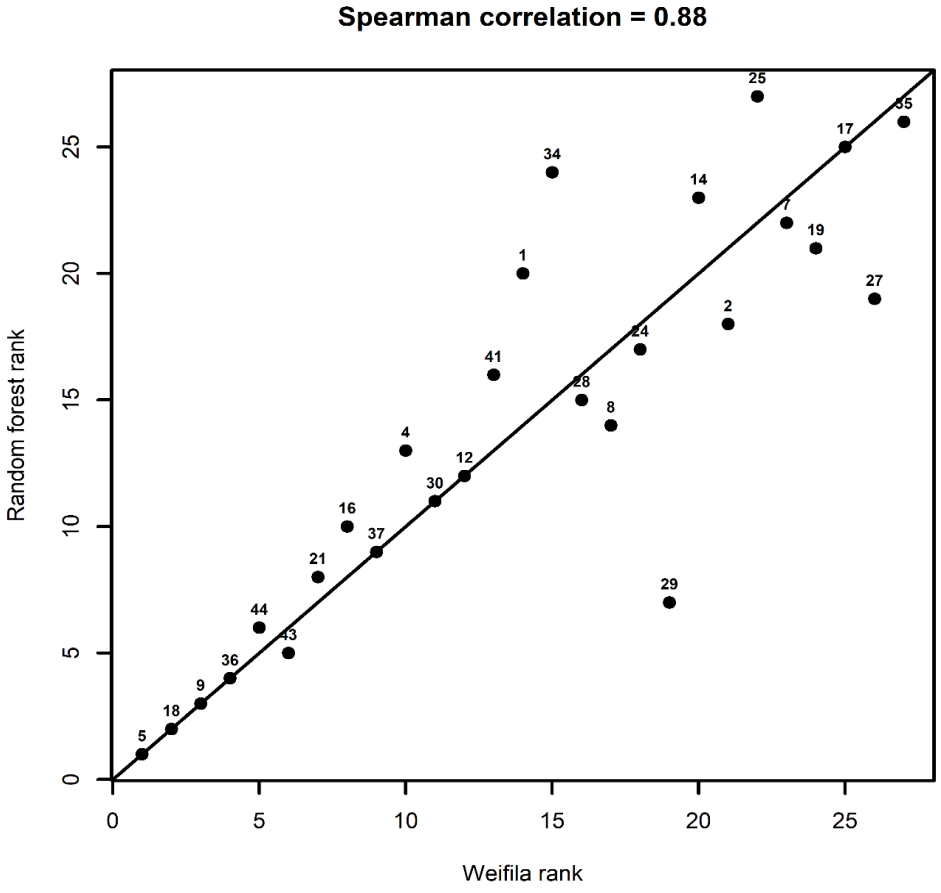
Appendix D: Prevalence, importance, RI and final ranking obtained for all selected 27 WPSFs using either the Weifila or the random forest method to assess their importance

For the Weifila approach, since L = 0.09 < R^2 = 0.19 < F = 0.56, **Equation 1** (Appendix A) was used to quantify importance of predictors $w(i)$

WPSF	Prevalence (%)	Weifila			Random Forest		
		Importance (%)	RI	Ranking	Importance (%)	RI	Ranking
1	38	2.4	91.8	**14**	1	37.1	**20**
2	61	0.9	56.1	**21**	0.7	44.8	**18**
4	49	2.6	126.2	**10**	1.3	64	**13**
5	43	8.6	372.1	**1**	12.1	522.3	**1**
7	24	1.4	33.2	**23**	1.3	32.6	**22**
8	60	1.3	77.7	**17**	0.9	51.6	**14**
9	15	17.5	262.1	**3**	26.5	396.7	**3**
12	42	2.7	110.6	**12**	1.7	72	**12**
14	41	1.6	63.8	**20**	0.7	30.3	**23**
16	37	4.1	150	**8**	2	74.1	**10**
17	29	0.6	18.2	**25**	0.5	15.8	**25**
18	27	13.5	362.5	**2**	16.3	436.9	**2**
19	62	0.4	26.6	**24**	0.6	35.8	**21**
21	28	5.7	162.3	**7**	3.4	95.3	**8**
24	32	2.3	73.9	**18**	1.4	45	**17**
25	37	1.5	55.9	**22**	0	0.1	**27**
27	75	0.1	8.8	**26**	0.6	43.6	**19**
28	49	1.7	82	**16**	1	49.4	**15**
29	22	3.3	72.9	**19**	5.2	113.3	**7**
30	30	3.8	115.7	**11**	2.4	73.1	**11**
34	57	1.5	86	**15**	0.5	29.4	**24**
35	71	0	1.1	**27**	0.1	9.3	**26**
36	43	5.6	237.3	**4**	6.8	288.8	**4**
37	24	6	143.5	**9**	3.2	75.6	**9**
41	53	2.1	110	**13**	0.9	46.4	**16**
43	42	4.5	191.1	**6**	5.6	235.4	**5**
44	52	4.1	215.8	**5**	3.2	169.6	**6**

Abbreviations: WPSFs, work-related psychosocial factors, RI, ranking index.

Appendix E: Scatter plot depicting the correlation between the two final rankings obtained for all 27 work-related psychosocial factors, using the Weifila approach and the random forest approach. Digits indicated the WPSF number.



Appendix F: Inter-correlations between the GHQ-28 subscales and the total scale

	GHQ-28 score	GHQ somatic symptoms	GHQ anxiety/insomnia	GHQ social disfunction	GHQ severe depression
GHQ-28 score	1				
GHQ somatic symptoms	0.85	1			
GHQ anxiety/insomnia	0.89	0.73	1		
GHQ social disfunction	0.78	0.58	0.55	1	
GHQ severe depression	0.76	0.47	0.58	0.50	1

Appendix G: List of the 10 priority WPSFs identified using the ranking index, for the four GHQ-28 subscales as well as for the total score

WPSF rank	**Somatic symptoms**	**Anxiety and insomnia**	**Social disfunction**	**Severe depression**	**Total GHQ-28**
1	44	5	5	9	5
2	5	36	18	43	18
3	18	18	9	18	9
4	21	43	28	37	36
5	8	9	44	36	44
6	36	34	30	29	43
7	12	21	4	1	21
8	16	16	17	5	16
9	43	41	14	24	37
10	9	8	16	12	4

References

1 Wallard H. Using Explained Variance Allocation to analyse Importance of Predictors. *16th ASMDA, Conference Proceeding* 2015;:1043-1054. Available from: http://www.asmda.es/images/1_W-Z_ASMDA2015_Proceedings.pdf.

2 Grömping U. Relative Importance for Linear Regression in *R* : The Package **relaimpo**. *Journal of Statistical Software* 2006;**17**. doi:10.18637/jss.v017.i01

3 Grömping U. Variable importance in regression models. *Wiley Interdisciplinary Reviews: Computational Statistics* 2015;**7**:137–52. doi:10.1002/wics.1346

4 Breiman L. Random Forests. *Machine Learning* 2001;**45**:5–32. doi:10.1023/A:1010933404324

5 Classification and Regression Trees. CRC Press. https://www.crcpress.com/Classification-and-Regression-Trees/Breiman-Friedman-Stone-Olshen/p/book/9780412048418 (accessed 2 Jan 2019).

6 Goldstein BA, Hubbard AE, Cutler A, *et al.* An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. *BMC Genet* 2010;**11**:49. doi:10.1186/1471-2156-11-49

7 Lebedev AV, Westman E, Van Westen GJP, *et al.* Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness. *Neuroimage Clin* 2014;**6**:115–25. doi:10.1016/j.nicl.2014.08.023

8 Hothorn T, Hornik K, Strobl C, *et al. party: A Laboratory for Recursive Partytioning*. 2018. https://CRAN.R-project.org/package=party (accessed 31 Jan 2019).

9 Hothorn T, Hornik K, Zeileis A. Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics* 2006;**15**:651–74. doi:10.1198/106186006X133933