

## **The 'double whammy' of low prevalence in clinical risk prediction**

Thomas R. Fanshawe

Nuffield Department of Primary Care Health Sciences

University of Oxford

Radcliffe Primary Care Building

Radcliffe Observatory Quarter

Woodstock Road

Oxford

OX2 6GG

[thomas.fanshawe@phc.ox.ac.uk](mailto:thomas.fanshawe@phc.ox.ac.uk)

Seena Fazel

Department of Psychiatry

University of Oxford

Warneford Hospital

Oxford

OX3 7JX

Word count: 2559

## **Background**

Worldwide, around 800,000 people die each year from suicide,[1] which is the leading cause of death in the UK in young adults.[2] Prediction modelling studies have attempted to incorporate demographic, clinical and other factors to identify high-risk individuals so that appropriate interventions can be offered.[3, 4]

This approach has a large literature but has not always been judged successful. In one review, all 35 suicide risk prediction studies assessed were classified as having high risk of bias or insufficient diagnostic accuracy, based on targets of 80% sensitivity and 50% specificity.[5] Others have written of a performance ‘glass ceiling’ in suicide prediction, and even that “risk categorization of individual patients has no role to play in preventing the suicide of psychiatric inpatients”.[6]

In spite of the mortality data quoted above, in most populations risk of death from suicide is low, usually below 1%. This has led to claims about performance of prediction models that appear counterintuitive, such as “suicide prediction models produce accurate overall classification models, but their accuracy of predicting a future event is near 0”, while noting that even in high-risk populations, the positive predictive value (PPV) of many prediction rules may be less than 1%.[3]

We describe two principal reasons why the nature of developing clinical prediction rules in low prevalence scenarios will almost invariably result in concerns about performance, on standards by which this is usually judged. Although we focus primarily on suicide prediction, these points apply more generally to other low prevalence clinical areas, some examples of which are also discussed.

### **1: The effect of low prevalence on sample size**

Low prevalence can have a prohibitive impact on sample size requirements because of the need to observe enough outcome events for model development, as also noted in diagnostic evaluation studies.[7] For example, consider a single risk factor that is present in half of individuals, in a population with outcome prevalence in those without this factor is 1%. To detect a relative risk (RR) of 2 with 90% power (5% significance level), requires a sample size of over 5,000 (equation 8 of [8]). More plausibly, if this risk factor occurs in a minority of individuals (say 10%), this sample size jumps to over 14,000. Figure 1 shows the pattern for larger values of the RR.

Developing prediction models also requires consideration of the effects of using multiple predictors during model selection, overfitting, and validation, which may increase the required sample size manyfold.[9] Although specific methods for determining adequate sample size when developing prediction models depend on the strength and nature of the associations between the included variables,[9] a good rule of thumb is to include at least ten outcome events for each risk factor examined, to prevent overfitting.[10] Multiplicity of risk factors can become a major issue when the number of predictors is at an extreme, as exemplified by a machine learning study in the US general population that examined 2,978 risk factors in a dataset containing 222 suicide events.[11] Another study based in a single university hospital used 272 risk factors in a dataset with 33 suicide events.[12] The large sample sizes required may make single-site studies infeasible in low prevalence scenarios.

### **2: The effect of low prevalence on predictive performance**

The second issue relates to the effect size required of the predictors that would result in levels of prognostic performance deemed acceptable in practice. The population prevalence ( $p$ , e.g. 1%) can be considered as the outcome probability if no risk factor information is available. As shown in the Supplementary Material, the relationship between the RR representing the effect of a risk factor (or the combined effect of a set of risk factors), the PPV, the sensitivity ( $S$ ) and  $p$  is then given by

$$RR = \frac{PPV - pS}{p(1 - S)}.$$

This allows the relationship between RR,  $S$  and PPV to be illustrated for given prevalence. This can be a useful concept because relative risk as an effect size measure is familiar across many research areas. For the 1% prevalence scenario, Figure 2 plots RR against  $S$  for varying PPV. For a PPV of 50%, the RR can never be less than 50, and in the range of sensitivities usually seen as acceptable, it will be higher still. Even for a PPV of 10%, which is higher than that reported for most suicide risk prediction models, the required RR would need to be well over 10.[13]

An alternative measure that is often used to summarise the relationship between prevalence and outcome probability calculated from a risk prediction model is the likelihood ratio. The positive likelihood ratio is defined as Sensitivity/(1-Specificity) and can be interpreted as the change in the odds of the outcome after using the prediction tool. To convert a prevalence 1% to a probability of 10%, 25% or 50% would require the prediction tool to correspond to a positive likelihood ratio of 11, 33 and 99 respectively. Even though it has been noted the likelihood ratio measures themselves may vary with outcome prevalence,[14] the conclusion is the same as that obtained from the RR, as likelihood ratios of this magnitude may appear unachievable.[15]

## Implications

The two issues described in this paper constitute a ‘double whammy’ of low prevalence in risk prediction studies. Effect sizes that are plausible for risk factors require investigations with large sample sizes to reach acceptable statistical power, but predictive performance often appears inadequate even if this sample size is met.

The first issue described may be ameliorated using linked databases of clinical populations, although even these may be insufficient for tools that target less populous subgroups. There is a potential conflict between the size of these databases and the nature of the variables collected: larger routine databases are more likely to contain broad demographic and clinical information than data concerning regular individual-level monitoring, such as symptom change, and are therefore unable to capture short-term changes in symptoms and behaviour that may be associated with increased risk.[16]

The second issue cannot be resolved by increasing sample size alone. Risk prediction using machine learning, often using national datasets, has become popular as an alternative,[17] but the effect of introducing this methodological complexity, as recommended by some authors,[18] appears likely to bring at best an incremental improvement in performance. One review suggested, albeit with uncertainty, that the best performing machine learning methods in suicide prediction might bring performance equivalent to an odds ratio of 12, much less than what would be required under the second condition outlined above.[19] Across a range of epidemiological research areas, relative risks and likelihood ratios of the required magnitude are rarely observed and cannot be realistically expected.[20] While these relative measures may help in developing conceptual understanding of

the general issues relating to low prevalence, we recommend that they be presented alongside absolute measures of risk when they are reported in particular studies.[21]

An alternative strategy is to use a more prevalent outcome. In suicide prediction, such an outcome is self harm, which represents a large burden of morbidity, or suicidal ideation. But the uncertainty here is whether risk stratification will lead to changes in clinical practice, if the outcome is not seen as clinically meaningful, and given current estimates of predictive accuracy for self harm outcomes.[22] Although suicide ideation is associated with death from suicide, this association is imperfect and the majority of patients with suicide ideation do not die from suicide.[23]

The considerations outlined in this paper suggest that there should more reasonable expectations about the achievable performance of prediction tools. In suicide risk assessment, there are at least two reasons to consider why a step-up in an individual's risk from 1% (population prevalence) to 5-10% (modelled: a modest absolute change but large relative change) may be informative when interpreted alongside additional clinical assessment.

Firstly, prediction rules may contain modifiable risk factors that could be targets for treatment, such as suicidal ideation, recent non-adherence to treatment, and comorbidities (for a review of risk factors, see [18]), which allows modifiable risk factors to be embedded into structured risk assessment to improve management and facilitate communication about risk in different healthcare settings. Secondly, the knowledge that an individual is at higher than average risk could be used to underscore safety planning to mitigate that risk.[24] This relatively low-cost intervention, a key component of which is to identify risk factors that may contribute to an individual having elevated risk, has shown promise in reducing suicidal behaviours.[25] However, it must also be acknowledged that a range of other suicide prevention interventions have been demonstrated to be effective, and some of these do not require a specific risk assessment to be made.[26] Some of these are service-level approaches.

## **Conclusion**

The prevalence of the outcome is a key consideration when planning and carrying out studies of clinical prediction rules. As this article has demonstrated, if the outcome prevalence is low, required sample sizes may become prohibitively large and the calculated risk of individuals identified as those at the highest risk may nevertheless remain quite low on the 0-100% scale in absolute terms.

In the field of suicide prediction, given the likely ceiling on the predictive performance of prediction rules, one implication of this is that prediction rules should not be used in isolation for clinical decision-making and allocation of interventions. Instead, their more appropriate role should be considered as an adjunct for decision-making, and the process of using tools in this way should be scrutinised in respect of their validation, translation into practice, and impact on patient outcomes to ensure that decisions are not harmful for the individual.[27] This requires assessment of outcomes, costs and consequences following the adoption of the prediction tool.[28, 29]

**Competing interests**

Both authors have previously published research relating to the development and validation of risk prediction tools, including a suicide risk prediction tool for individuals with severe mental illness (OxMIS).

**Funding**

This research was supported by the National Institute for Health Research Applied Research Collaboration Oxford and Thames Valley at Oxford Health NHS Foundation Trust. TF also receives funding from the NIHR Community Healthcare MedTech and In Vitro Diagnostics Co-operative at Oxford Health NHS Foundation Trust (MIC-2016-018). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. SF is additionally funded by the Wellcome Trust (202836/Z/16/Z).

## References

1. World Health Organisation. Suicide [accessed 27/01/2021]. Available from: <https://www.who.int/news-room/fact-sheets/detail/suicide>.
2. Office for National Statistics. Deaths registered in England and Wales: 2019 [accessed 27/01/2021]. Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathsregistrationsummarytables/2019>.
3. Belsher BE, Smolenski DJ, Pruitt LD, Bush NE, Beech EH, Workman DE, et al. Prediction models for suicide attempts and deaths: a systematic review and simulation. *JAMA Psychiatry*. 2019;76(6):642-51.
4. Chan MKY, Bhatti H, Meader N, Stockton S, Evans J, O'Connor RC, et al. Predicting suicide following self-harm: systematic review of risk factors and risk scales. *Brit J Psychiat*. 2016;209(4):277-83.
5. Runeson B, Odeberg J, Pettersson A, Edbom T, Jildevik Adamsson I, Waern M. Instruments for the assessment of suicide risk: A systematic review evaluating the certainty of the evidence. *PLOS ONE*. 2017;12(7):e0180292.
6. Large M, Ryan C, Nielssen O. The validity and utility of risk assessment for inpatient suicide. *Australas Psychiatry*. 2011;19(6):507-12.
7. Holtman GA, Berger MY, Burger H, Deeks JJ, Donner-Banzhoff N, Fanshawe TR, et al. Development of practical recommendations for diagnostic accuracy studies in low-prevalence situations. *J Clin Epidemiol*. 2019;114:38-48.
8. Woodward M. Formulae for sample size, power and minimum detectable relative risk in medical studies. *J R Stat Soc Ser D*. 1992;41(2):185-96.
9. Riley RD, Snell KI, Ensor J, Burke DL, Harrell Jr FE, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med*. 2019;38(7):1276-96.
10. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996;49(12):1373-9.
11. García de la Garza Á, Blanco C, Olsson M, Wall MM. Identification of suicide attempt risk factors in a national US survey using machine learning. *JAMA Psychiatry*. 2021.
12. Krupinski M, Fischer A, Grohmann R, Engel R, Hollweg M, Möller H-J. Risk factors for suicides of inpatients with depressive psychoses. *Eur Arch Psy Clin N*. 1998;248(3):141-7.
13. Large M, Smith G, Sharma S, Nielssen O, Singh S. Systematic review and meta-analysis of the clinical factors associated with the suicide of psychiatric in-patients. *Acta Psychiatr Scand*. 2011;124(1):18-9.
14. Brenner H, Gefeller OJ. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Stat Med*. 1997;16(9):981-91.
15. Grimes DA, Schulz KF. Refining clinical diagnosis with likelihood ratios. *Lancet*. 2005;365(9469):1500-5.
16. Ballard ED, Gilbert JR, Wusinich C, Zarate CA. New methods for assessing rapid changes in suicide risk. *Front Psychiatry*. 2021;12(31).
17. Gradus JL, Rosellini AJ, Horváth-Puhó E, Street AE, Galatzer-Levy I, Jiang T, et al. 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.
18. Franklin JC, Ribeiro JD, Fox KR, Bentley KH, Kleiman EM, Huang X, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. *Psychol Bull*. 2017;143(2):187-232.
19. Corke M, Mullin K, Angel-Scott H, Xia S, Large M. Meta-analysis of the strength of exploratory suicide prediction models; from clinicians to computers. *BJPsych Open*. 2021;7(1):e26.

20. Ioannidis JPA, Panagiotou OA. Comparison of effect sizes associated with biomarkers reported in highly cited individual articles and in subsequent meta-analyses. *JAMA*. 2011;305(21):2200-10.
21. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ*. 2010;340:c869.
22. Carter G, Milner A, McGill K, Pirkis J, Kapur N, Spittal MJ. Predicting suicidal behaviours using clinical instruments: systematic review and meta-analysis of positive predictive values for risk scales. *Br J Psychiatry*. 2017;210(6):387-95.
23. Hubers A, Moaddine S, Peersmann S, Stijnen T, Van Duijn E, Van der Mast R, et al. Suicidal ideation and subsequent completed suicide in both psychiatric and non-psychiatric populations: a meta-analysis. *Epidemiol Psych Sci*. 2018;27(2):186.
24. Stanley B, Brown GK. Safety planning intervention: a brief intervention to mitigate suicide risk. *Cogn Behav Pract*. 2012;19(2):256-64.
25. Stanley B, Brown GK, Brenner LA, Galfalvy HC, Currier GW, Knox KL, et al. Comparison of the safety planning intervention with follow-up vs usual care of suicidal patients treated in the emergency department. *JAMA Psychiat*. 2018;75(9):894-900.
26. Hofstra E, Van Nieuwenhuizen C, Bakker M, Özgül D, Elfeddali I, de Jong SJ, et al. Effectiveness of suicide prevention interventions: a systematic review and meta-analysis. *Gen Hosp Psychiatry*. 2020;63:127-40.
27. Whiting D, Fazel S. How accurate are suicide risk prediction models? Asking the right questions for clinical practice. *Evid Based Ment Health*. 2019;22(3):125.
28. Kessler RC, Bossarte RM, Luedtke A, Zaslavsky AM, Zubizarreta JR. Suicide prediction models: a critical review of recent research with recommendations for the way forward. *Molec Psychiat*. 2020;25(1):168-79.
29. Douglas T, Pugh J, Singh I, Savulescu J, Fazel S. Risk assessment tools in criminal justice and forensic psychiatry: The need for better data. *Eur Psychiatry*. 2017;42:134-7.

## Figure legends

**Figure 1.** Total sample size required to detect a relative risk of the given size, with 90% power at the 5% significance level, if the prevalence of the outcome in the comparison group is 1% and a certain percentage of the population have the risk factor (5%, 10%, 20% or 50%, as shown by the separate lines).

**Figure 2.** Relative risk as a function of sensitivity, in a population with outcome prevalence of 1%, for different values of the positive predictive value (10%, 20% or 50%, as shown by the separate lines)



Figure 1

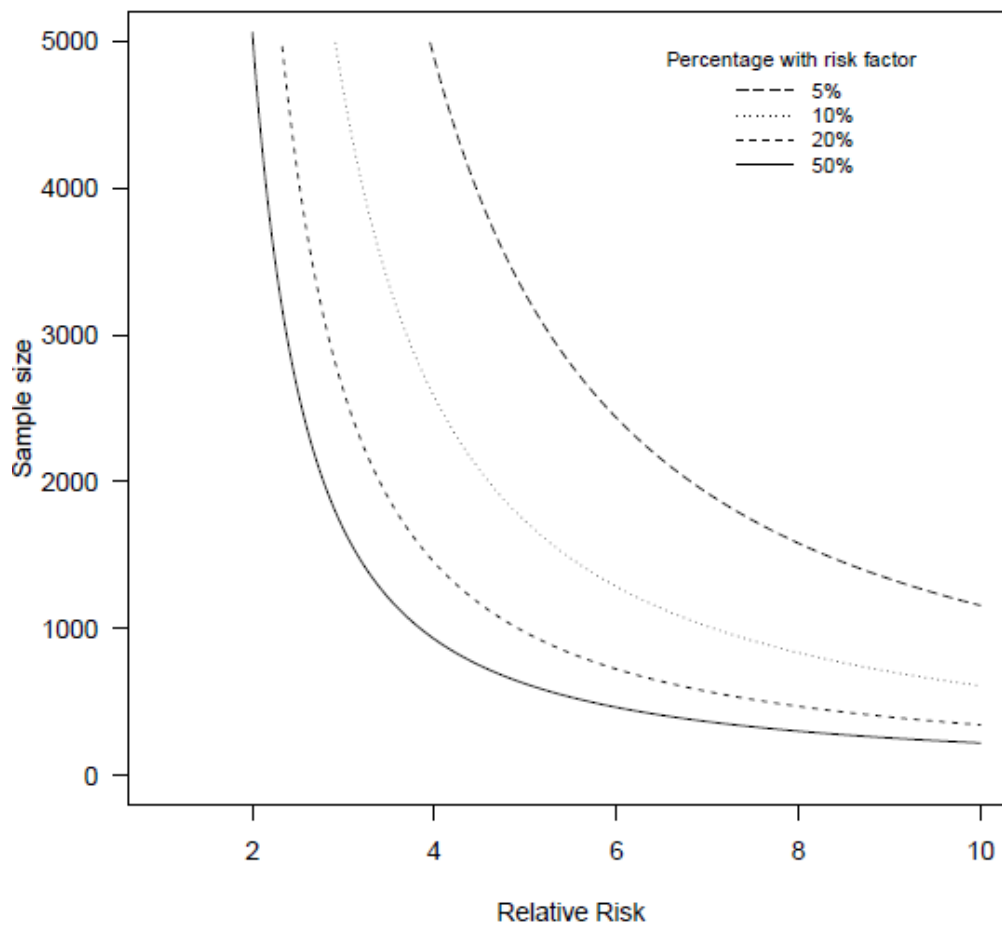
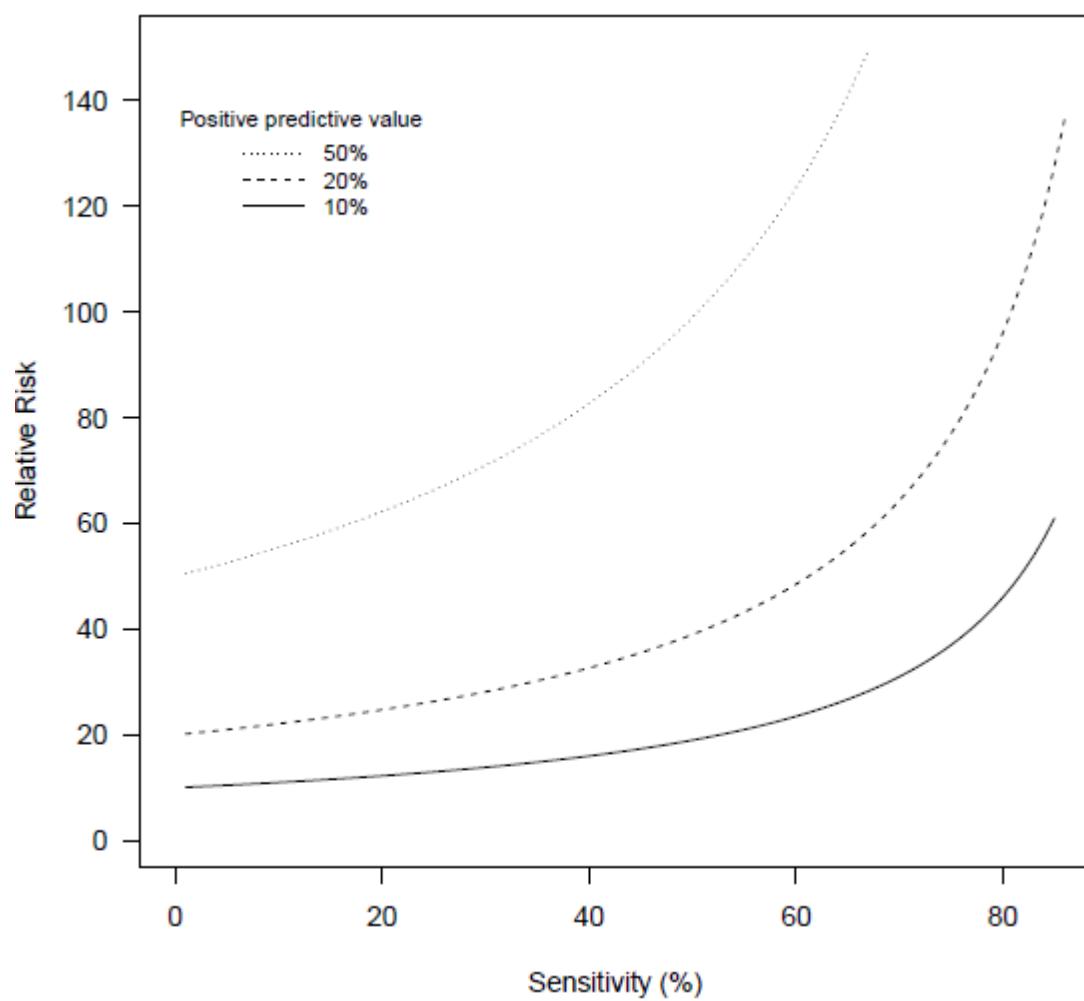


Figure 2



## Supplementary Material

### Derivation of equation in Section 2

For a given risk factor and outcome event, we use the following notation:

	Outcome positive	Outcome negative	Total
Risk factor positive	a	b	a+b
Risk factor negative	c	d	c+d
Total	a+c	b+d	n=a+b+c+d

Additionally write  $p$  for the prevalence of the outcome,  $RR$  for the relative risk of the risk factor in relation to the outcome event,  $S$  for its sensitivity, and  $PPV$  and  $NPV$  for the positive and negative predictive values, respectively.

By the definitions of these quantities, we have

$$p = \frac{a + c}{n}$$

$$S = \frac{a}{a + c} = \frac{a}{np}$$

$$PPV = \frac{a}{a + b}$$

$$NPV = \frac{d}{c + d}$$

$$RR = \frac{a/(a + b)}{c/(c + d)} = \frac{PPV}{1 - NPV}$$

It follows that

$$1 - NPV = \frac{c}{c + d}$$

$$= \frac{n((a + c)/n) - a}{n - a((a + b)/a)}$$

$$= \frac{np - a}{n - a/PPV}$$

Thus

$$RR = PPV \times \frac{n - a/PPV}{np - a}$$

$$= \frac{nPPV - a}{np - a}$$

$$= \frac{PPV - pS}{p(1 - S)}$$