Published in final edited form as:
Evid Based Ment Health. 2019 August 01; 22(3): 125–128. doi:10.1136/ebmental-2019-300102.

How accurate are suicide risk prediction models? Asking the right questions for clinical practice

Daniel Whiting[iD], **Seena Fazel**
Department of Psychiatry, University of Oxford, Oxford, UK

Abstract

Prediction models assist in stratifying and quantifying an individual's risk of developing a particular adverse outcome, and are widely used in cardiovascular and cancer medicine. Whether these approaches are accurate in predicting self-harm and suicide has been questioned. We searched for systematic reviews in the suicide risk assessment field, and identified three recent reviews that have examined current tools and models derived using machine learning approaches. In this clinical review, we present a critical appraisal of these reviews, and highlight three major limitations that are shared between them. First, structured tools are not compared with unstructured assessments routine in clinical practice. Second, they do not sufficiently consider a range of performance measures, including negative predictive value and calibration. Third, the potential role of these models as clinical adjuncts is not taken into consideration. We conclude by presenting the view that the current role of prediction models for self-harm and suicide is currently not known, and discuss some methodological issues and implications of some machine learning and other analytic techniques for clinical utility.

Introduction

Providing information on prognosis is routine in modern medicine, and guides clinical decisions about further investigations and treatments. Such predictions are typically made by clinical judgement, which may or may not be informed by evidence about risk factors. However, they have been increasingly combined with statistical models and tools for a stratified, more precise approach to treatment. Prognostic information also provides patients and carers with information about their future health and function in order to help them plan their lives and care accordingly. One common example is the widespread use of cardiovascular risk calculators, such as the Framingham or QRISK scores, that can help guide whether statin therapy is considered. Other areas, such as cancer medicine, frequently

Daniel Whiting: 0000-0001-5323-364X

Correspondence to: Professor Seena Fazel, Department of Psychiatry, University of Oxford, Oxford OX3 7JX, UK; seena.fazel@psych.ox.ac.uk.

Contributors DW and SF conceived, wrote and edited the article.

Disclaimer The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health and Social Care.

Competing interests SF is coauthor of a 2019 paper on the development and validation of a suicide risk assessment tool in severe mental illness.

Provenance and peer review Not commissioned; externally peer reviewed.

Europe PMC Funders Author Manuscripts

use prognostic tools to inform specific treatment choices. New technologies, including the availability of large datasets, have led to a flood of new prediction models, which will be one of most significant impacts of information technology on the future of healthcare delivery.[1]

In psychiatry, suicide is one of the few adverse outcomes that informs clinical practice at all levels from referral and assessment to treatment. But it is difficult to predict. Prevention at both the population level and targeted at high risk groups is recommended, but there are contrasting views on whether prediction models or tools might assist in this process. Their predictive accuracy and clinical utility is questioned, and in England, for example, while assessing risk is a core part of practice and an explicit feature of decision-making about whether to detain someone under the Mental Health Act, making decisions based on predicting the risk of suicide or self-harm using a tool is not recommended by national guidelines.[2] Instead, clinicians should typically undertake an unstructured clinical assessment of factors they deem relevant and focus on clinical and psychosocial needs.

Methods

We searched PubMed for the 3 years up to end April 2019 using the keywords ‘meta-analysis’, ‘systematic’, ‘assess*’, ‘predict*’, ‘suicid*’ and ‘self-harm’, and selected three recent systematic reviews that have specifically considered the predictive accuracy of tools and models used to predict self-harm and suicide, and included information on a range of approaches, from adapted scales to models derived by traditional statistical and machine learning methods, which were examined in psychiatric and general hospital as well as community settings. Runeson *et al*[3] and Carter *et al*[4] examined a variety of scales used to predict risk (either by design or adaptation), while Belsher *et al*[5] focused on whether models derived by newer data techniques, such as machine learning, have led to predictions that are accurate enough to be clinically useful (table 1).

This clinical review will present a critical appraisal of these reviews, in which we discuss the need to compare tool performance with current clinical practice, consider a range of performance measures and also address specific methodological aspects of model development that have been overlooked in these reviews. For these reasons, we think that the broad conclusion drawn by these reviews, that risk prediction for suicide is not possible nor clinically useful, is premature. Rather we propose that a more evidence-based and balanced interpretation is that the potential clinical utility of suicide risk prediction currently is unknown. We also consider how future research might address this.

Presentation

Carter *et al*[4] undertook a meta-analysis of three types of instrument used to predict suicide death or self-harm: psychological scales, biological tests and ‘third-generation’ scales derived from statistical models. The review was framed to examine their clinical usefulness, with a focus on performance based on one metric, positive predictive value (PPV). Inclusion criteria defined a risk assessment tool as any scale to which a cut-off score was applied to designate risk status. The original purpose of each scale is not reported. However, of 36

different psychological scales included, 6 are in fact rating scales for intense affect, depression or anxiety, 2 are personality inventories and 1 is a drug misuse screen.

For suicide, this review estimated that the pooled random-effects estimate PPV for psychological instruments was 3.7% (95% CI 2.5% to 5.4%) and for biological measures 14.5% (95% CI 9.4% to 21.7%). For self-harm, the pooled PPV for psychological instruments was 27.5% (95% CI 22.8% to 32.7%) and for biological measures 14.7% (95% CI 6.3% to 30.8%). For third-generation scales, which combined self-harm studies with suicide, the pooled PPV was 38.7% (95% CI 26.9% to 51.9%), although most of the contributing primary studies had a high risk of bias. This review concluded that no instrument was sufficiently accurate to determine intervention, and suggested alternatives to using risk assessment for allocating future healthcare: adopting a needs-based approach to reduce exposure to modifiable risk factors, or allocating interventions for subpopulations (such as diagnostic groups) or to unselected clinical populations (such as offering psychotherapeutic interventions to all presenting to hospital following self-harm).

A second meta-analysis[3] reviewed the sensitivity and specificity of 15 different instruments for suicide and suicide attempt, which included those tools intended for prediction and also originally developed for other purposes. Sensitivity and specificity varied widely between tools, and none, either in individual studies or for five tools where meta-analysis was possible, achieved the arbitrarily chosen benchmark of 80% sensitivity and 50% specificity. This benchmark would mean clinically that one in five individuals with an adverse outcome would be missed by any tool (ie, false negatives), and that one in two individuals deemed high risk would not develop the outcome (ie, false positives). Although not included in their consideration of utility, negative predictive value (NPV, the proportion of those identified as low risk who do not develop the outcome) was reported in their supplementary material, and ranged from 76% to 100%.

More recently, Belsher *et al*[5] set out to evaluate models specifically developed for the prediction of suicidal behaviours and whether advances in modelling had improved algorithms ‘sufficiently to render their predictions actionable’. They searched for investigations that longitudinally evaluated models and that included both development and testing stages. This systematic review focused on two performance metrics—an overall measure of discrimination, the area under the receiver operating characteristic (AUC) curve, which can be interpreted as the probability of correctly classifying pairs of subjects with and without the outcome and the PPV. The AUCs for models predicting suicide mortality ranged from 0.59 to 0.86, and PPVs from <0.1% to 19%. For models that predicted suicide attempts, AUCs ranged from 0.71 to 0.93, and PPVs from 0% to 78%. Sensitivity was also reported (ranging from 6% to 94% for suicide mortality and 11% to 96% for suicide attempt), but key performance metrics including the NPV or model calibration (comparing observed with expected probabilities) were not. The latter is particularly important as a model may distinguish well between individuals with and without the outcome (discrimination), but poorly estimate the probability of events in a target population (calibration). The authors concluded that, although overall discrimination was good across most models, PPVs were mostly ‘extremely low’ and so these models currently offer limited practical utility.

Europe PMC Funders Author Manuscripts

Limitations

We outline four limitations in these reviews, which render their conclusions questionable.

First, findings were not compared with current clinical practice, where risk assessment is routine, inconsistent and might perform better or worse than these tools. The reviews discuss how the rarity of the outcome places a ceiling on positive predictive power, however this challenge applies equally to unstructured clinical judgement. Any interpretation of the performance of prediction models should therefore also discuss how current unstructured approaches perform. These clinical approaches are unlikely to be accurate. To illustrate this, another review pooled data from studies reporting the longitudinal relationship between specific risk factors (including those typically used in clinical practice) and suicide outcomes, and showed these risk factors performed little better than chance, whether treated individually or as categories (eg, weighted AUC for prior self-injurious thoughts and behaviours was 0.61 (SE 0.02) for suicide attempt, and 0.59 (SE 0.03) for suicide death).[6] Another example that demonstrates the current performance of clinical judgement is a national survey of psychiatric services in England and Wales, where over three-quarters of individuals who died by suicide during 10 years were judged low or no risk at their last clinical contact.[7] This equates to a sensitivity of <25% for clinical judgement of an increased risk of suicide. One interpretation is that risk assessment is not possible; another is that it is done poorly, with scope for improvement by supporting clinical decision-making with even modestly performing models. Important questions for future research are how statistical models compare to unstructured clinical judgement, or lead to incremental benefit when used to support such judgement, and how their statistical performance can guide the nature of their clinical application.

A second problem with these reviews is that they examine the predictive accuracy of tools without reference how they would be linked to a clinical decision. The review by Belsher *et al* uses the term ‘actionable’ as the accuracy threshold to determine utility of prediction models, but importantly this depends on the subsequent intervention, and without clarification of this, interpretation of the findings is not possible. If used to identify who to assess more fully (as suggested in some primary studies), or improve stratification to a non-harmful intervention by helping to target those who would derive the greatest absolute reduction in risk, a high false positive rate may be acceptable. For example, based on Framingham prediction scores for cardiovascular events, clinical guidelines deem 7.5% a sufficiently high probability of a future outcome to make this a threshold at which to consider statin therapy. In other words, of 100 people who might be prescribed statins, >90 would not experience a cardiovascular event in the subsequent 10 years even without a statin.

Third, PPV is the performance measure on which two of these reviews focus their conclusions about clinical applicability, but the value of NPV should not be ignored. Belsher *et al* note that NPV will be high with a rare outcome, and Runeson *et al* briefly discusses the anomaly that NPV may be artificially high if also using a small number of predictors that are themselves rare, regardless of their individual relationship to the outcome. However, rather than discounting NPVs, these are reasons to consider a range of performance measures rather than one in isolation, and derive models using meaningful predictors with transparent

reporting of their relationship to the outcome. This is particularly the case where there are many predictors that are related to suicide outcomes—here a high NPV is important. Some suicide models have very high NPVs,[8] and harnessing this aspect of performance to support clinicians to consistently, transparently and accurately judge low risk may have high clinical utility through preserving resources.[9] By safely screening out individuals at lower risk of suicide, services can focus their limited resources by further assessing and/or treating those at elevated risks.

Finally, to consider the quality of studies that develop predictive algorithms and models, particular methodological characteristics need to be addressed.[10] All three reviews use the QUADAS-2 tool to evaluate quality.[11] However, this scale was designed for diagnostic accuracy studies, and is less applicable to prediction models developed with newer methods and large datasets. New rating scales should be used, such as PROBAST,[10] which have been developed for prediction models. This is a particular limitation of the review by Belsher *et al*, which focuses on models that have used machine learning approaches. Two important criteria—testing and reporting model calibration, and reporting sample size as events per candidate variable—were present in only one of the 10 included studies from 2009 to 2018 predicting suicide mortality (table 2). This questions this particular review's conclusion that quality of the primary studies was high. Furthermore, techniques such as machine learning pose distinct questions when considering clinical translation that should feature in such discussion (table 3).

Clinical Implications

For clinical practice, one consistent finding from these reviews is that prediction of suicide is difficult and associated with uncertainty. It is important that this is acknowledged by clinicians and services, and discussed openly with patients and carers. Nevertheless, we have tried to show that the extent to which these difficulties will prevent any helpful application of tools in clinical practice has been overstated.

The debate over using prediction models for suicide and selfharm leads to the wider clinical question of whether a stratified medicine approach to preventing suicide should be abandoned altogether—that would diverge psychiatry from much of the rest of medicine.[12] While needs-based approaches and universal prevention strategies have been proposed,[13] the current reality for all mental health services is that finite resources need targeted allocation. Some judgement of risk inevitably contributes to this, such as determining which patients with severe depression in primary care need referral to specialist mental health services,[14] and in clinical practice the separation between 'assessment' and 'prediction', endorsed by Carter *et al*, is likely to be an abstract concept. One alternative suggested by Carter *et al*, to offer all those who present with self-harm a psychological intervention, is not currently feasible, and so the clinical challenge remains of needing to assess risk and allocate intervention, for which the responsibility typically falls to clinicians' judgement alone.

It has also been argued that the process of stratifying risk detracts from undertaking a holistic, therapeutic assessment of needs.[15] These two do not have to be mutually exclusive, and it is possible to consider the situation where a tool acts as an adjunct or aid for clinical

decision-making that can improve efficiency and consistency,[16] and anchor assessments in an evidence base, thus giving clinicians greater confidence and time to focus on developing an individualised treatment plan, importantly shifting the focus away from lengthy risk assessments and on to risk management. This will form part of the process of translating advances in data science to clinical benefit.

What next in research?

Prognostic model research across medicine is too weighted towards the development of new models, of which very few are taken through a comprehensive evaluation within a clinical setting.[17] This remains pertinent with the increasing accessibility of electronic health records and use of machine learning techniques. Indeed, it is noticeable how few of the algorithms in the review by Belsher *et al* seek to produce an output that would allow independent validation or clinical pilot work. The emphasis needs to shift far closer to the clinical setting to address questions regarding practical applicability.

To examine the clinical utility of suicide prediction tools, future evaluations should test performance compared with current unstructured approaches, and when used as adjuncts to support assessment and decision-making in a clearly defined place in clinical pathways. The strengths and limitations of a model, and its performance on different measures of accuracy, need to be explicitly considered when determining clinical role. Continuing to appraise free-floating model performance without this framing substantially limits clinical relevance. Measures of reclassification, determining how often a tool's rating correctly differs from an unstructured clinical judgement if categorical ratings are used, can also be informative. Evaluations should additionally consider the calibration of a model for a target population,[18] and studies of clinical impact will need to consider a range of different outcome measures (rather than just predictive performance), as well as the various contextual factors that affect clinical implementation and use.

Conclusion

Whether prediction models and risk assessment tools can be applied to suicide prevention remains an open question. While the primary studies included in three recent systematic reviews do not provide evidence for clinical implementation, the reviews themselves are limited and overstate their conclusions because they do not compare models with current approaches or consider the value of high NPVs. Rather than continuing to develop new models in isolation, future work needs to move towards real-world clinical evaluations that examine the incremental benefits of using these tools to support clinical decision-making rather than replace it.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding

SF is funded by the Wellcome Trust (202836/Z/16/Z). DW is funded by the National Institute for Health Research (NIHR Doctoral Research Fellowship, DRF-2018-11-ST2-069).

Europe PMC Funders Author Manuscripts

References

1. Health Education England. The Topol Review. Preparing the healthcare workforce to deliver the digital future. Health Education England; 2019.
2. National Institute for Health and Care Excellence. Self-harm in over 8s: long-term management [CG 133]. NICE; 2011.
3. Runeson B, Odeberg J, Pettersson A, et al. Instruments for the assessment of suicide risk: a systematic review evaluating the certainty of the evidence. PLoS One. 2017; 12:e0180292. [PubMed: 28723978]
4. Carter G, Milner A, McGill K, et al. Predicting suicidal behaviours using clinical instruments: systematic review and meta-analysis of positive predictive values for risk scales. Br J Psychiatry. 2017; 210:387–95. [PubMed: 28302700]
5. Belsher BE, Smolenski DJ, Pruitt LD, et al. Prediction models for suicide attempts and deaths: a systematic review and simulation. JAMA Psychiatry. 2019
6. Franklin JC, Ribeiro JD, Fox KR, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. Psychol Bull. 2017; 143:187–232. [PubMed: 27841450]
7. The National Confidential Inquiry into Suicide and Safety in Mental Health Annual Report. University of Manchester; 2018.
8. Fazel S, Wolf A, Larsson H, et al. The prediction of suicide in severe mental illness: development and validation of a clinical prediction rule (OxMIS). Transl Psychiatry. 2019; 9:98. [PubMed: 30804323]
9. Bolton JM, Gunnell D, Turecki G. Suicide risk assessment and intervention in people with mental illness. BMJ. 2015; 351
10. Wolff RF, Moons KGM, Riley RD, et al. PROBAST Group†. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019; 170:51–8. [PubMed: 30596875]
11. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011; 155:529–36. [PubMed: 22007046]
12. The Academy of Medical Sciences (AMS). Realising the potential of stratified medicine. AMS; 2019.
13. Large MM. The role of prediction in suicide prevention. Dialogues Clin Neurosci. 2018; 20:197–205. [PubMed: 30581289]
14. National Institute for Health and Care Excellence. Depression in adults: recognition and management [CG90]]. NICE; 2009.
15. Large MM, Ryan CJ, Carter G, et al. Can we usefully stratify patients according to suicide risk? BMJ. 2017; 359
16. Berman NC, Stark A, Cooperman A, et al. Effect of patient and therapist factors on suicide risk assessment. Death Stud. 2015; 39:433–41. [PubMed: 25674940]
17. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013; 10:e1001381. [PubMed: 23393430]
18. Fazel S, Wolf A. Selecting a risk assessment tool to use in practice:a 10-point guide. Evid Based Ment Health. 2018; 21:41–3. [PubMed: 29269440]

Europe PMC Funders Author Manuscripts

Table 1

Selected elements and findings of three recent systematic reviews of approaches to suicide and self-harm prediction

Review	Types of approach included	Primary study populations	Performance measures	Findings and conclusions
Runeson *et al*[3]	Psychological rating scales with risk cut-off applied Prediction tools (unweighted variables)	Psychiatric patients (inpatient and outpatient) Individuals presenting to emergency settings Primary care patients	Sensitivity Specificity (NPV/PPV in supplement)	None achieved predefined accuracy threshold (80% sensitivity, 50% specificity). No support for use. Unclear whether may improve prediction as complement to clinical impression.
Carter *et al*[4]	Biological measures with risk cut-off applied Psychological rating scales with risk cut-off applied Prediction tools (unweighted and weighted variables)	Psychiatric patients (inpatient and outpatient) Individuals presenting to emergency settings Military veterans Prisoners	PPV LR/CUI * summarised	Combined pooled PPV 26.3% for self-harm and 5.5% for suicide. No individual instrument or pooled subgroup with accuracy suitable to allocate treatment.
Belsher *et al*[5]	Prediction models derived by various methods (including machine learning)	Psychiatric patients (inpatient and outpatient) Individuals presenting to emergency settings Primary care patients Military populations General population	AUC Accuracy Sensitivity PPV	Good overall classification, but low PPV. Would result in high false-positive and considerable false-negative rates if used in isolation. At present limited practical utility.

* Performance metrics primarily applied to diagnostic and screening tests.

AUC, area under the receiver operating characteristic curve; CUI, clinical utility index; LR, likelihood ratio; NPV, negative predictive value; PPV, positive predictive value.

Europe PMC Funders Author Manuscripts

Table 2

Assessment of methodological quality of studies from 2009 to 2018 reviewed by Belsher *et al* using selected items of PROBAST tool[10]

Study	Events per variable	Handling of missing data	Calibration plot or table comparing predicted vs observed outcome probabilities
Amini *et al*, 2016	NR	Expectation maximisation	NR
Barak-Corren *et al*, 2017	NR	Complete-case analysis	NR
Choi *et al*, 2018	NR	NR	NR
DelPozo-Banos *et al*, 2018	NR	NR	NR
Ilgen *et al*, 2009	NR	NR	NR
Kessler *et al*, 2015	NR	Nearest neighbour, multiple and rational imputation	NR
Kessler *et al*, 2017	NR	Nearest neighbour and rational imputation	NR
Kessler *et al*, 2017b	NR	Complete-case analysis	NR
McCarthy *et al*, 2015	NR	NR	NR
Simon *et al*, 2018	NR	NR	Tabulated but not reported graphically

NR, not reported. References in online supplementary file 1.

Europe PMC Funders Author Manuscripts

Table 3
Comparison of regression and machine learning approaches to clinical prediction

Regression methods	Machine learning methods
Informed by assumptions, background knowledge and theory.	Exploratory, data-driven, automatically learns from data.
Typically use a small number of variables to predict probability of an outcome.	May be more suited to handling a large number of predictors in data with high signal-to-noise ratio.
Mainly linear effect of variables on outcome.	More flexible, captures non-linear associations and interactions between variables, strategies required to reduce overfitting.
Provide clinically informative relationships between variables and outcome, allows, for example, consideration of counterfactuals.	Limited clinical interpretability, 'black-box' algorithms may lack face validity for clinicians, especially if large number of unintuitive predictors.
Results often simply presented for end-user, for example, conversion to a score.	Transparent presentation of results difficult.
Can undertake model updating for use in populations with different baseline risk.	Testing calibration and updating to new baseline risk difficult for many models.