LETTER TO THE EDITOR

Improper study design precludes valid effect estimates in important suicide prevention research

Abstract

The observational study design for estimating accuracy of diagnostic tests for suicide risk in clinical work is not ideal, due to the effects of directed suicide prevention to the high-risk group. This is an example of the confounding by indication and protopathic bias, which lead to misinterpretation of the accuracy terms sensitivity and specificity. The simple arithmetic mechanism presented here, forces the conclusion that the accuracy estimates sensitivity and specificity, applied in open prospective trials of surmised diagnostic tests for suicide risk, cannot be normally interpreted as accuracy estimators. Further, the related concept “prediction of suicide” is shown to be fundamentally illogical and should not be used in the present context. All these statements reveal a far-reaching problem within the suicide prevention research: Conclusions regarding the usefulness of diagnostic tests of suicide risk in the vast previous research since decades deserve reinterpretation. Diagnostic tests per se, can possess highly qualitative properties in estimating important suicidological aspects regarding a patient, but, when studied in an open design study, it cannot be demonstrated. This problem concerns rating scales and any biological and psychological tests in medicine, when confounding factors, for example, suicide prevention, influence the outcome because of the test result per se.

There is a serious lack of valid methods for detecting patients who are at risk of suicide and to whom preventive measures should effectively be directed (Chan et al., 2016; Large et al., 2016; Runeson et al., 2017). In establishing valid diagnostic instruments for suicide risk, their accuracy is estimated. Most often in medical research on diagnostic tests, the outcome (disease or not) is already known at the test occasion, but not so often in suicide prevention research in which the outcome (suicide or not) appears after a varying period of follow-up. In most trials of surmised diagnostic instruments for estimating suicide risk, when based on clinical, psycho/social and, or, demographic data, the assessments are made by clinicians. If not, they are informed of the test results. Consequently, high-risk patients will be subject to directed suicide preventive measures in higher degree than low-risk patients during the follow-up period. This is an example of *confounding by indication* (Joseph, Mehrabadi, & Lisonkova, 2014)—or *protopathic bias* (Faillie, 2015) depending on the causality of the events—how systematic errors occur when estimation of the predictive ability of a diagnostic test is combined with preventive measures directed to the predictive outcome. This type of confounding is not specific for suicide prevention research; it can be seen for example in research

on diagnostic instruments for breast cancer. Screening for breast cancer by means of mammography leads to diagnostic and treatment efforts that by themselves confound the unadulterated effect of the mammography by itself on mortality. In this and similar cases, the overall mortality in breast cancer in the screened and comparison groups is compared, since mortality is usually of primary interest irrespective of the presence of confounding factor(s). Confounding by inclusion affects also blood pressure measurements, and the effect of the white coat when they are made at the hospital.

Usually, the estimation of the value of diagnostic tests in medicine is based on the 2×2 table invented by Pearson and Filon (1898), described by Stigler (2002) and Altman and Bland (1994). The number of patients with positive (high risk) and negative (low risk) test results are allocated in the 2×2 table according to whether they commit suicide or not during a fixed or variable follow-up period (Figure 1). It is obvious that the number in cell A, in the figure true positives, decreases by one and that in cell B, false positives, increases by one for each patient that is saved from death by suicide in the high-risk group.

Thus, in statistical terms, successful suicide prevention leads to a reduction of true positives, that is, a decrease of suicide victims among the high-risk patients, and an increase of false positives, that is, growth of the number of survivors in the high-risk patients. It follows that with successful prevention, the sensitivity to suicide approaches zero (i.e., no true positives and maximum number of false positives), which, in fact, is the goal of suicide prevention in health care—the *vision of suicide zero*. However, the opposite opinion has been prevailing since the 1970s—and still is—that the proportion of true positives should be as high as possible, that is, sensitivity as high as possible. The other way around—low sensitivity—is considered as an error by Diggory (1974, p. 60) and unwanted by Runeson et al. (2017) to name examples. To sum up, in open naturalistic studies in the current context, the sensitivity parameter is no longer an estimate of the accuracy of the test, rather an indicator of the effectiveness of suicide prevention.

Further, the parameter specificity, the proportion of false positives, is heavily dependent on unforeseen physical or psychosocial factors as discussed before (Thorell et al., 2014) and should not be used even in blind studies. Also, the Positive Predictive Value (PPV) is dependent on the numbers in cells A and B (Figure 1) and thus becomes invalid in this situation. Consequently, the content in the 2×2 table, cells A and B, contains information that is invalid and has led to misinterpretations and invalid conclusions since the 1970s when contingency tables began to be used in studies of suicide risk (Buglass & McCulloch, 1970; Rosen, 1976) therefore valid for conclusions. The

		Outcome:	
Test result:		SUICIDE	NOT SUICIDE
POSITIVE	TRUE A Suicides in high-risk group Target for directed prevention based on test and other reasons	↓	FALSE B Survivors among high-risk group ↑
NEGATIVE	FALSE C Suicides among low-risk group Prevention for other reasons No test based prevention		TRUE D Survivors among low-risk group
		Sensitivity = $A/(A+C)$ Share of the high-risk group among suicides ↓	Specificity = $D/(D+B)$ Share of suicides among low-risk persons ↓
		PPV = $A/(A+B)$ Share of suicides in high-risk persons ↓	NPV = $D/(C+D)$ Share of survivors in low-risk group

FIGURE 1 The more effective the directed suicide prevention, the smaller the number of suicides among high-risk patients (true positives) and the larger the numbers of survivors among high-risk patients (false positives) due to intervention based on a positive test result. PPV is Positive Predictive Value and NPV Negative Predictive Value. Arrows designate increase or decrease of numbers in cells when one or more patients avoid suicide in the high-risk group due to intervention based on a positive test result

NPV is, though, influenced by other factors such as the “general rate” of suicides in psychiatric care.

Now, what is the problem really? The accuracy terms per se are not the problem, nor are, necessarily, the diagnostic tests. The crucial problem is the design of studies: The follow-up period allows, and even demands, external impact—suicide prevention—thus introducing a strong source for bias to the outcome. Consequently, the commonly used prospective open, naturalistic design in suicide prevention research, in which the clinicians are aware of the test results, is not ideal, and offers results that often are misinterpreted: A high sensitivity is usually interpreted as if the diagnostic test possesses high accuracy for suicide risk—while it is likely that the suicide prevention in such studies is ineffective in the high-risk group and little can be said about the test accuracy. The other way around, a low sensitivity usually is not welcome (Runeson et al., 2017)—while it is likely that a large proportion of the patients judged to be at risk have been saved from suicide, maybe partly because the test is highly accurate.

Apparently, new approaches are necessary to be developed for evaluating diagnostic tests for suicide risk. A controlled experimental, blind design with suicide/survival as the binary outcome for two groups of patients with positive or negative test result would be a first choice. However, although such a design seems to be a clean experimental approach, the design is open for confounding. Patients will still receive preventive measures on other grounds than the test result. This will happen when a risk of suicide is clinically detected, for example, through a perceived serious verbal or nonverbal suicidal message. Thus, a, by the name “blind,” design would not guarantee complete blindness in real-life trials. However, the most serious argument against this blind, controlled experimental design is the ethics. When the researcher has a grounded belief that the instrument really can differentiate between subjects, it will be unethical to prevent patients with a positive test result from clinically motivated suicide preventive measures and treatment.

When a blind experimental design cannot properly be applied, the alternative is an open naturalistic approach. However, as the present paper shows, the independent effect on mortality or other relevant

variables of diagnostic methods for risks of suicide must be seriously questioned.

But, there would be a tenable way to estimate the relative contribution of a diagnostic instrument for suicide risk, if the degree, or better impact, of applied suicide preventive efforts could be estimated and be used as a covariate in the statistical analysis. As far as known, such type of estimation does not exist today. Thus, there is a great need to develop such methods. Consequently, the ethically most defensible studies of the contribution of diagnostic instruments for suicide risk, or better, vulnerability to suicide, seems to be of the naturalistic type. But new statistical approaches must be developed. A conclusion of the present discussion is that ethical and adequate methods in dealing with the confounding effects of treatments in research on the detection of risks of suicide do not seem to exist. However, when a valid method for quantification or at least a definition of the inner mechanisms of the disease behind the suicide has become established, the accuracy of the diagnostic test should be directed to the prediction of the existence of that mechanism instead of suicide/survival. Such knowledge has not yet been published.

The parameter sensitivity may be justified in blind studies in which the clinicians are unaware of the test results or unaware of its possible relationship to the risk of suicide—provided that other observable variables are not informative vis-à-vis the outcome of the diagnostic test.

This is often true in evaluations of biological and psychological tests for suicide risk if preventive efforts are not directed specifically to the high-risk group. An example is a study of a psychophysiological test by Thorell and Wahlin (2018). To some degree this may be true also in evaluation of tests based on register data. Some diagnostic tests contain information on gender, marital status, employment, social class, substance abuse, physical illness, and psychiatric disorders (see, e.g., Qin, Agerbo, & Mortensen, 2003). To the extent that clinical risk assessors comprehend similar evaluations for the register data as the constructors of the very test, the study is not blind.

In addition, when confounding by indication prevails, any parameter estimate, related to the sensitivity and specificity (the cells A

and B), such as the odds ratio or relative risk or in other ways relates outcome to the diagnostic test result, such as regression methods, will be biased.

Large et al. (2016) pointed out that naturalistic studies within the current topic have a weakness and write: "Successful interventions provided to high-risk patients in the primary studies may have the effect of reducing the odds of suicide in that group." Further, they state that "The extent of this effect cannot be estimated without studies that directly investigate the effectiveness of providing increased resource allocation or enhanced clinical surveillance to high-risk patients." (Large et al., 2016). Thus, it seems that the problem focused on in the present letter is not unknown, but not enough emphasized, why it is found necessary here to stress its substantial consequences on the development of diagnostic tools for risk of suicide. The suggestion of measures of estimating the impact of resource allocation or enhanced clinical surveillance to high-risk patients is a way to improve future development (Large et al., 2016). Also, Runeson et al. (2017) stated that

ideally, the diagnostic accuracy should be tested in controlled trials where patients randomly are assigned to risk assessment or no risk assessment. Such a design is not an option since patients who are potentially suicidal need to be evaluated and those identified as being at high risk are offered other interventions than their peers. This may in itself influence future risk of suicide, thus altering the instrument's predictive ability.

However, the consequences of this reasoning were not explicitly drawn for the prerequisites of their review study that used a study-inclusion criterion of at least 80% sensitivity (Runeson et al., 2017). The motivation of the present letter is obvious: to point out that the consequences of confounding by indication must be accentuated much stronger than before in order to correct the 40-year-old research topic in suicide prevention and to develop other designs and methods for evaluating the validity of diagnostic tests of suicide risk.

Linked to this problem is that the often-used concept of *prediction of suicide* is unsuitable in the present context and should not be used. Even if used in a pure statistical way, it should be avoided for educational clarity. Prediction of suicide, based on individual characteristics, is meaningless, since suicide can be triggered or prevented by any of the immense number of unexpected factors—bad news, good news, a fence, a meeting with someone, that is, everything that hinder or facilitate a method or occasion for the suicidal act. In addition, predicting suicide is of no clinical interest—when literally predicted, it is *fait accompli*: what could be done for a patient who is predicted to commit suicide, whatever action is taken? Prediction implies predestination that for suicide obviously does not exist since it can be prevented.

If diagnostic instruments for suicide risk point out what can be considered an expression of a disease, the present discussion should involve the statements by Wilson & Jungner (1968). Referring to these authors, it can be said that although suicide is an important

health problem with a battery of accepted preventive measures, facilities are still lacking for diagnosis, also true for latent or early symptomatic stages or an acceptable test. Further, the natural history of the condition, including development from latent to declared disease, is not adequately understood, and thus, an agreed policy on whom to treat as risk patients is still lacking, and therefore, the cost of case-finding cannot be proved to be economically balanced in relation to possible expenditure on medical care as a whole. The basic cause why the Wilson and Jungner (1968) criteria cannot be applied to the present health problem is that suicide is a behaviour that can be prevented or triggered by non-medical factors (fence, harassing), not a disease. Although the case-finding is in many clinics a continuing process, but still, it seems to be a "once and for all" event.

However, today, a new objective test for suicide risk detection and for possible diagnoses of a neuropsychological dysfunction behind the vulnerability to suicide in depression has been studied since 1987. It is an objective test of electrodermal reactivity (EDOR) in an experiment on habituation of the electrodermal response (Thorell, 1987). It seems possible that the suicide risk-related orienting hyporeactivity expresses a previously unknown basic neuropsychological disease for which the Wilson and Jungner criteria may apply. It is, however, out of scope of the present paper to develop this issue further.

1 | CONCLUSION

The open design, which is used in naturalistic studies of surmised diagnostic tests for suicide risk in clinical work, leads to a strong bias affecting the outcome, that is, confounding by indication, and to misunderstanding of the actual meaning of accuracy parameters, constituting a serious obstacle to the development of this urgent section of suicide prevention research. Another problem to take into consideration is the usual dichotomization of a continuous observational variable when defining a dichotomous goal variable such as suicide/survival (Brenner & Gefeller, 1997). Ideally, the observed test data should have a bimodal distribution. Further, there is a risk that diagnostic tests and scientific work, past and current, of high quality has not become published and even lost for the future due to the finding of low sensitivity. For example, one of the recent reviews of diagnostic tools for suicide (Runeson et al., 2017) excluded all studies with results of lower sensitivity than 80%. Perhaps the most interesting studies were not considered—most interesting since many patients at risk likely were saved due to the combination of possibly an accurate diagnostic test and effective suicide prevention. The value of clinical diagnostic tests has been seriously questioned since 1970s (Diggory, 1974) and up until today in recent reviews (Chan et al., 2016; Large et al., 2016; Runeson et al., 2017). It is necessary to make researchers and clinicians aware of the problem behind and to establish new approaches for enabling sound estimates of the values of diagnostic tests of suicide risk in order to improve the urgently important suicide preventive work.

DECLARATION OF INTEREST STATEMENT

Lars-Håkan Thorell is the funder, shareholder, and employee of the medical-technical research company Emotra AB, Ltd, Gothenburg, Sweden. Karl Wahlin is the owner of the statistical consulting company Karl Wahlin AB, Stockholm, Sweden. Jonas Ranstam is the owner of the statistical consulting company Mdas AB, Rotfruktsgatan 12B, Ystad, Sweden.

ACKNOWLEDGEMENTS

None.

Keywords

confounding by indication, naturalistic studies, sensitivity, suicide prevention research, trials of diagnostic tests for suicide risk

ORCID

Lars-Håkan Thorell <https://orcid.org/0000-0002-6843-3975>

Karl Wahlin <https://orcid.org/0000-0002-2485-0419>

Jonas Ranstam <https://orcid.org/0000-0002-8287-7273>

Lars-Håkan Thorell^{1,2}

Karl Wahlin³

Jonas Ranstam⁴

¹Research and Development, Emotra AB (Ltd), Gothenburg, Sweden

²Department of Clinical and Experimental Medicine, Linköping University, Linköping, Sweden

³Karl Wahlin AB, Stockholm, Sweden

⁴Mdas AB, Ystad, Sweden

Correspondence

Lars-Håkan Thorell, Research and Development, Emotra AB (Ltd), Göteborgsvägen 88, 43363 Sävedalen, Gothenburg, Sweden, Repslagaregatan 40, 58222 Linköping, Sweden.
Email: lars@emotra.se

REFERENCES

- Altman, D. G., & Bland, J. M. (1994). Diagnostic tests. 1: Sensitivity and specificity. *BMJ*, 308, 1552.
- Brenner, H., & Gefeller, O. (1997). Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Statistics in Medicine*, 16, 981–991.
- Buglass, D., & McCulloch, J. W. (1970). Further suicidal behaviour: The development and validation of predictive scales. *The British Journal of Psychiatry*, 116, 483–491. PMID: 5460558. <https://doi.org/10.1192/bjp.116.534.483>
- Chan, M. K., Bhatti, H., Meader, N., Stockton, S., Evans, J., O'Connor, R. C., ... Kendall, T. (2016). Predicting suicide following self-harm: Systematic review of risk factors and risk scales. *The British Journal of Psychiatry*, 209, 277–283. <https://doi.org/10.1192/bjp.bp.115.170050>
- Diggory, J. (1974). Predicting suicide: Will-o-the-wisp or reasonable challenge? In A. T. Beck, H. L. P. Resnik, & D. Lettieri (Eds.), *The prediction of suicide* (pp. 59–70). Bowie: Charles Press.
- Faillie, J. L. (2015). Indication bias or protopathic bias? *British Journal of Clinical Pharmacology*, 80, 779–780. <https://doi.org/10.1111/bcp.12705>
- Joseph, K. S., Mehrabadi, A., & Lisonkova, S. (2014). Confounding by indication and related concepts. *Current Epidemiology Reports*, 1, 1–8. <https://doi.org/10.1007/s40471-013-0004-y>
- Large, M., Kaneson, M., Myles, N., Myles, H., Gunaratne, P., & Ryan, C. (2016). Meta-analysis of longitudinal cohort studies of suicide risk assessment among psychiatric patients: Heterogeneity in results and lack of improvement over time. *PLoS ONE*, 11, e0156322. <https://doi.org/10.1371/journal.pone.0156322>
- Pearson, K., & Filon, L. N. G. (1898). Mathematical contributions to the theory of evolution IV. On the probable errors of frequency constants and on the influence of random selection on variation and correlation. *Philosophical Transactions of the Royal Society of London*, 191, 229–311. Reprinted in *Karl Pearson's Early Statistical Papers*, Cambridge: Cambridge University Press, 1956, pp. 179–261. <https://doi.org/10.1098/rsta.1898.0007>
- Qin, P., Agerbo, E., & Mortensen, P. B. (2003). Suicide risk in relation to socioeconomic, demographic, psychiatric, and familial factors: A national register-based study of all suicides in Denmark, 1981–1997. *The American Journal of Psychiatry*, 160, 765–772. <https://doi.org/10.1176/appi.ajp.160.4.765>
- Rosen, D. H. (1976). The serious suicide attempt. Five-year follow-up study of 886 patients. *JAMA*, 235, 2105–2109. PMID: 946536. <https://doi.org/10.1001/jama.1976.03260450017021>
- Runeson, B., Odeberg, J., Pettersson, A., Edbom, T., Jildevik Adamsson, I., & Waern, M. (2017). Instruments for the assessment of suicide risk: A systematic review evaluating the certainty of the evidence. *PLoS ONE* 2017, 12, e0180292. <https://doi.org/10.1371/journal.pone.0180292>
- Stigler, S. (2002). The missing early history of contingency tables. *Annales de la faculté des sciences de Toulouse 6e série*, 11, 563–573. <https://doi.org/10.5802/afst.1039>
- Thorell, L. H. (1987). Electrodermal activity in suicidal and nonsuicidal depressive patients and in matched healthy subjects. *Acta Psychiatrica Scandinavica*, 76, 420–430. <https://doi.org/10.1111/j.1600-0447.1987.tb05627.x>
- Thorell, L.H., Wahlin, K. (2018). Electrodermal hyporeactivity as a marker for vulnerability to suicide in depression. *EC Psychology and Psychiatry Open Access* 7, <https://www.econicon.com/ecpp/pdf/ECPP-07-00344.pdf>
- Thorell, L. H., Wolfersdorf, M., Straub, R., Steyer, J., Hodgkinson, S., Kaschka, W. P., ... Wahlin, K. (2014). A paradox in suicide statistics in estimating specificity of tests for suicide—Reply to Mushquash and co-workers and Culver. *Journal of Psychiatric Research*, 54, 142–143. <https://doi.org/10.1016/j.jpsychires.2014.03.015>
- Wilson, J.M.G., Jungner, G. (1968). Principles and practice of screening for disease. Public health papers No. 34, World Health Organization, Geneva.