Author/s:
Woodford, R;Spittal, MJ;Milner, A;McGill, K;Kapur, N;Pirkis, J;Mitchell, A;Carter, G

Title:
Accuracy of Clinician Predictions of Future Self-Harm: A Systematic Review and Meta-Analysis of Predictive Studies

Date:
2019-02-01

Citation:
Woodford, R., Spittal, M. J., Milner, A., McGill, K., Kapur, N., Pirkis, J., Mitchell, A. & Carter, G. (2019). Accuracy of Clinician Predictions of Future Self-Harm: A Systematic Review and Meta-Analysis of Predictive Studies. Suicide and Life Threatening Behavior, 49 (1), pp.23-40. https://doi.org/10.1111/sltb.12395.

Persistent Link:
https://hdl.handle.net/11343/293541

Article type : Original Article

Title

Accuracy of clinician predictions of future self-harm: a systematic review and meta-analysis of predictive studies

Authors

Rachel Woodford [a]

Matthew J. Spittal [b]

Allison Milner [b]

Katie McGill [a e]

Navneet Kapur [c]

Jane Pirkis [b]

Alex Mitchell [d]

Gregory Carter [e]

Affiliations

[a] School of Medicine and Public Health, Faculty of Health and Medicine, University of Newcastle, NSW, Australia

[b] Melbourne School of Population and Global Health, University of Melbourne, Parkville, VIC, Australia

[c] Centre for Suicide Prevention, University of Manchester, UK
[d] Cancer Studies & Molecular Medicine, University of Leicester, UK
[e] Centre for Brain and Mental Health Research, University of Newcastle, NSW, Australia

Corresponding Author
Professor Gregory Carter
Email: Gregory.carter@newcastle.edu.au
Phone: +61 (0) 240144926
Fax: +61 (0) 240144933
Mailing Address: Locked Bag #7, Hunter Region Mail Centre, 2301. NSW, Australia

Abstract

Assessment of a patient after hospital-treated self-harm or psychiatric hospitalisation often includes a "risk assessment", resulting in a classification of high risk versus low risk for a future episode of self-harm. Through systematic review and a series of meta-analyses looking at unassisted clinician risk classification (8 studies; n= 22,499), we found pooled estimates for sensitivity 0.31 (95% CI: 0.18-0.50), specificity 0.85 (0.75-0.92), Positive Predictive Value (PPV) 0.22 (0.21-0.23) and Negative Predictive Value (NPV) 0.89 (0.86-0.92). Clinician classification was too inaccurate to be clinically useful. After-care should therefore be allocated on the basis of a needs rather than risk assessment.

Hospital-treated self-harm is very common in the general hospital Emergency Department. Self-harm is the commonly used term in the UK (National Institute for Health and Care Excellence, 2011), usually defined as *"self-poisoning or injury, irrespective of the apparent purpose of the act"*; especially in reference to hospital-treated self-harm, which is predominately self-poisoning with a minority of self injury (Carroll, Metcalfe, & Gunnell, 2014). Hospital-treated self-harm is the strongest independent risk factor for subsequent suicide (Carroll, Metcalfe, & Gunnell, 2014). However, repetition of self-harm is a much more common outcome than suicide; with pooled estimates for 12 month prevalence in

predominately western countries being 15-16% for repetition of self-harm and 1-1.6% for suicide (Owens, Horrocks, & House, 2002; Carroll, Metcalfe, & Gunnell, 2014). After discharge from adult psychiatric hospital inpatient care (for any reason), more than 6% of people are admitted to the general hospital for an episode of self-harm within 12 months, with one third of these episodes occurring in the month after discharge (Gunnell, Hawton, Ho, Evans, O'Connor et al, 2008).

Even in these two "high risk" clinical populations (hospital-treated self-harm and discharged psychiatric inpatients), suicide is uncommon in absolute terms; and it has been recognised for more than 60 years that low-frequency disorders like suicide are impossible to predict with a clinically useful positive predictive value (PPV). This is because the very low prevalence of suicide imposes a ceiling on the PPV (Rosen, 1954); and this ceiling is independent of the predictive method used (Carter, Milner, McGill, Pirkis, Kapur et al, 2017). In the general hospital and the psychiatric hospital setting, a typical goal is to distinguish between patients at "high" and "low" risk for suicide in order to allocate after-care, although this approach has been criticised: "*...it is simply not possible to predict suicide in an individual patient, and any attempt to subdivide patients into high-risk and low-risk categories is at best unhelpful and at worst will prevent provision of useful and needed psychiatric care*" (Ryan & Large, 2013). A recent systematic review of 53 risk assessments from 37 studies demonstrated this more precisely, finding the pooled estimates for sensitivity (56%) and PPV (5.5%) for the classification of high-risk of future suicide (Large, Kaneson, Myles, Myles, Gunaratne et al, 2016). A second systematic review estimated a pooled PPV (5.5%) (Carter, Milner, McGill, Pirkis, Kapur et al, 2017) and a third review reported a range of PPV (1.3 to 16.7%) (Chan, Bhatti, Meader, Stockton, Evans et al, 2016), for a "high risk" of suicide classification. There have been seven recent systematic reviews that when considered collectively have demonstrated that any individual risk factor, instrument or scale used to classify patients into a high risk group for later suicide may do so modestly better than chance, however around half of all suicides come from the "low risk" classification, whilst very few of the patients in the "high risk" classification die by suicide (Carter, Page, Large, Hetrick, Milner et al, 2016; Large, Kaneson, Myles, Myles, Gunaratne et al, 2016; Chan, Bhatti, Meader, Stockton, Evans et al, 2016; Chapman, Mullin, Ryan, Kuffel, Nielssen et al, 2015; Hubers, Moaddine, Peersmann, Stijnen, van Duijn et al, 2016; Franklin, Ribeiro, Fox, Bentley, Kleiman et al, 2017; Ribeiro, Franklin, Fox, Bentley, Kleiman et al, 2016). None of these systematic reviews included studies of unassisted clinician prediction.

However, since self-harm in clinical populations is much more common than suicide, the question arises: "Can clinicians predict future self-harm with sufficient accuracy to be clinically useful?"

The state of clinical risk assessment in current practice for hospital-treated self-harm is difficult to know. In the UK there have been standards established for the assessment and care of self-harm patients by the Royal College of Psychiatrists (Royal College of Psychiatrists, 2010) and by clinical practice guidelines published by the National Institute for Health and Care Excellence (NICE) (National Institute for Health and Care Excellence, 2011), which suggest that any form of risk assessment to establish a "high risk" classification of patients, by clinician or validated instruments, in order to allocate treatment is not useful. However, *"Evidence also exists of widespread failure by NHS services to comply with the NICE Guideline on Self Harm"*(Royal College of Psychiatrists, 2010). A stratified sample of 32 UK hospitals found that risk classification was still being done; most commonly using "unvalidated locally developed proformas" (41% of emergency departments and 69% of mental health services) in the prediction of future self-harm, with others using a variety of standardised scales, clinician judgement or some combination of methods (Quinlivan, Cooper, Steeg, Davies, Hawton et al, 2014). In other parts of the world there may be greater reliance on clinician judgement as the usual method for determining risk of suicide or self-harm.

Unassisted (or partly assisted) clinical judgement is a form of risk assessment whereby clinicians stratify patients into various risk groupings (most commonly high vs low risk); and there are current recommendations from the USA that favour the use of a clinician risk assessment to classify patients into risk groups (high vs. low) in order to allocate treatment. "*The main procedure used by clinicians to determine whether an individual may be at risk of suicidal behaviors is the suicide risk assessment (SRA). The purpose of the SRA is to identify risk and protective factors that then provide the data for the formulation of suicide risk. The suicide risk formulation (SRF) assigns a level of suicide risk that ideally leads to triage and treatment deemed appropriate for that level of risk.*" (Silverman & Berman, 2014); or "*a comprehensive suicide risk assessment is used to collect detailed information about a patient's suicide risk (e.g., risk and protective factors), to detect the possibility of imminent risk, and to inform treatment decision"* recommended by the Suicide Prevention Resource Center (SPRC); or step 4 *"Determine Risk Level/Intervention (Determine risk. Choose appropriate intervention to address and reduce risk*)" and step 5 *"Document (Assessment of*

risk, rationale, intervention, and follow-up)" of the Suicide Assessment Five-step Evaluation and Triage (SAFE-T), which was developed in the US by the Substance Abuse and Mental Health Services Administration (SAMHSA) to assist "*clinicians in conducting a suicide assessment to identify risk factors and protective factors, conduct a suicide inquiry, determine risk level and potential interventions, and document a treatment plan"* (Suicide Prevention Resource Center, 2015).

The aim of the current systematic review and meta-analysis was to quantify the accuracy of unstructured clinician assessments of future risk of self-harm, in hospital-treated clinical populations with relatively high rates of subsequent self-harm. We calculated estimates of the key accuracy statistics for this task, especially the positive and negative predictive values. An understanding of the absolute degree of accuracy (or inaccuracy) of clinician judgement to determine risk stratification, will inform us about the usefulness of clinician prediction when positive (e.g., for the allocation of clinical after-care to prevent future self-harm events) and when negative (e.g., regarding the "safe to discharge home" decision).

Methodology

The study was prospectively registered on the PROSPERO database with registration number 2015:CRD42015015781(Woodford, Carter, McGill, Milner, Pirkis et al, 2015). We used the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) (Liberati, Altman, Tetzlaff, Mulrow, Gotzsche et al, 2009) to report the identified studies and the meta-analyses; and QUADAS rating system to assess risk of bias in diagnostic accuracy studies (Whiting, Rutjes, Westwood, Mallett, Deeks et al, 2011).

Searches were conducted of medical and scientific databases, including Medline, PsychInfo, EMBASE, CINHAL, Web of Science and Scopus, inclusive of all papers from database inception until January 2016. A secondary search of the papers identified in another systematic review regarding the accuracy of standardised and validated predictive instruments was also undertaken to identify any incidental clinician predictions (Carter, Milner, McGill, Pirkis, Kapur et al, 2017). Further papers were obtained by searching the reference lists of each paper retrieved; use of the 'find similar' functions for seminal papers in PubMed and Web of Science; and screening the reference lists of books dealing with the topic of clinician judgement. A search of the grey literature was conducted in January 2016 with review of the first 100 sources identified by the search engines Google and Google Scholar through combinations of the keywords.

Key terms used for the search included “self-harm” OR “suicide*”, together with their suggested synonyms, and “predict*” OR “forecast” AND “clinician” OR “clinical judgement” OR “clinical decision making”.

Inclusion and exclusion criteria

Eligible studies included those with original data on clinician prediction of future self-harm, or suicide attempt. These were required to be cohort studies of at least 10 participants, and include outcomes after any follow-up period. Study populations could include any non-paediatric population (≥12 years), with treatment settings including psychiatric hospital inpatients for any reason, and Emergency Department presentations or general hospital admissions for self-harm. Any studies reporting subgroups of the above populations were included, and data was pooled from studies that separated estimates on a basis of gender. Any studies that used “locally developed structured proformas” or validated instruments to determine risk assessment were excluded. Only studies published in English were included. Titles and abstracts recovered in the search were screened for study suitability and full text copies of papers were retrieved that possibly dealt with the review topic. Duplicated papers were removed at this point.

The retrieved papers were screened independently by two different reviewers (RW, GLC) and excluded or included on the basis of the above criteria. Final papers included a quantified assessment of clinician risk (high/medium/low) or (high/low).

Study Characteristics

Key study characteristics were extracted from the selected papers and presented in tabular form: author, year, clinical setting, country, clinician rater, gender of population and risk stratification.

Data extraction for accuracy statistics

Data from the final studies was pooled into high vs low risk as judged by clinicians. In papers that compared the rating accuracy of a number of clinician groups, additional sets of data were extracted for each clinician group. If a ‘medium or moderate’ risk judgement was made, this was combined with the ‘low’ assessment of risk to obtain two groups (high vs low risk); in order to achieve the combination most consistent with the proportions (high vs low risk) used in studies that only reported a 2-level classification of risk stratification. One study had combined the ‘moderate’ with the ‘high’ risk group for reporting (Cooper, Kapur, & Mackway-Jones, 2007); one study had combined the ‘moderate’ with ‘low’ risk group for reporting (Kapur, Cooper, Rodway, Kelly, Guthrie et al, 2005); and one study had an

additional stratification of 'high' and 'very high'(Phillips, Stargatt, & Brown, 2012), which we combined into 'high' risk for all analyses. Data were extracted from the studies directly into a 2 x 2 table, and any missing values were calculated from the data available.

Statistical Analyses

Accuracy statistics, Sensitivity, Specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV) and clinical utility as determined by Likelihood Ratio positive (LR+) and Likelihood Ratio negative (LR-) for self-harm outcomes for each individual study were calculated from the 2 x 2 tables using an online calculator, Diagnostic and AGreement Statistics (DAGStat) (http://www.biostats.com.au/DAG_Stat/#Top). A LR+ score in the range of 5-10 was considered to be a moderate increase in the likelihood of the predicted outcome, whilst a LR- score in the range of 0.1-0.2 was considered to be a moderate decrease in the likelihood of the predicted outcome, and hence potentially clinically useful. For each clinician prediction we also calculated the Clinical Utility of a positive (CUI+) (positive = high risk) or negative (CUI-) (negative = low risk) decision (see www.clinicalutility.co.uk) (Mitchell, 2011). CUI scores were considered in qualitative grades: excellent ≥ 0.81, good 0.80-0.64, satisfactory 0.63-0.49 and poor utility < 0.49 (Mitchell, 2011).

We calculated pooled estimates of sensitivity, specificity, PPV and NPV. The general form of our models was to estimate either the pooled sensitivity and specificity jointly (or the pooled PPV and the NPV jointly) using the binomial-normal model (Stijnen, Hamza, & Ozdemir, 2010). This is a random effects logistic regression model with an unstructured covariance matrix to allow a correlation between the study-specific estimates of sensitivity and specificity (or PPV and NPV). We estimated the pooled parameters on the logit scale and back-transformed them to proportions for presentation. Of the studies we examined, all but one followed individuals over time. The remaining study counted episodes (Cooper, Kapur, & Mackway-Jones, 2007), not persons and as such, we conducted a secondary analysis excluding this study from the pooled results.

Heterogeneity was assessed through visual inspection of the forest plots and with the I^2 statistic, which provides an estimate of inconsistency across studies. The meta-analyses of diagnostic values differ from the meta-analyses of intervention studies in several key ways. First, heterogeneity is to be expected; second, random effects models are used to estimate this heterogeneity (Macaskill, Gatsonis, Deek, Harbord, & Takwoingi, 2010); and third, the I^2 statistic overestimates heterogeneity in meta-analyses of diagnostic or predictive tests (Bossuyt, Davenport, Deeks, Hyde, Leeflang et al, 2013). Publication bias was assessed

using a modified version of Egger's test (Egger, Smith, Schneider, & Minder, 1997), which uses linear regression but is based on the efficient score and its variance (Harbord, Egger, & Sterne, 2006).

Risk of bias rating

Each clinician evaluation was rated independently by two reviewers (RW, GLC) for risk of bias using an adapted form of the QUADAS-2 (QUality Assessment of Diagnostic Accuracy Studies - version 2) rating instrument (Whiting, Rutjes, Westwood, Mallett, Deeks et al, 2011). This tool assessed risk of bias in four domains: patient selection (2 items: participant selection (random or consecutive) and exclusions less than 15% of population), index test (1 items: blinding to outcome), reference standard (2 items: classification of outcomes and blindness of rating), and flow and timing (3 items: duration of follow-up one year or less, same outcome measurement for all, drop-out less than 15%). For each of these rating points, the risk of bias was assessed as low, high or unclear. Any discrepancies in quality assessment were discussed and resolved by the two reviewers.

Results

The search yielded (n=5) five peer-reviewed published papers (Erdman, Greist, Gustafson, Taves, & Klein, 1987; Kapur, Cooper, Rodway, Kelly, Guthrie et al, 2005; Cooper, Kapur, & Mackway-Jones, 2007; Murphy, Kapur, Webb, & Cooper, 2011; Phillips, Stargatt, & Brown, 2012) and (n=1) one thesis from the grey literature (Naydock, 2015), which can be seen in the PRISMA diagram in Figure 1, for a total of (k= 8) eight clinician evaluations, seen in Table 1. Five evaluations were conducted in the Emergency Department of a general hospital, one in the Emergency Department of a veterans' hospital, one in an adult psychiatric hospital, and one in an adolescent psychiatric hospital. Five evaluations were done in UK populations, two in the US and one in Australia. A variety of clinicians conducted the assessments including Mental Health and Emergency Department staff. Seven evaluations predicted any individual having one or more future episodes with the unit of analysis as the individual; and one with the unit of analysis being events not individuals. Five evaluations used self-harm as the outcome, one used self-harm or suicide death and two used suicide attempt. Follow-up duration was 12 months for the general hospital populations, 24 months for the veterans' hospital population, and one month and 3 months for the psychiatric hospital populations. The range of the proportion of patients allocated to the high risk category for seven evaluations was (7.6-19.9%) and for one evaluation that used events rather than individuals and combined medium and high to produce a high risk group (65.8%) (Cooper, Kapur, & Mackway-Jones, 2007).

PUT FIGURE 1 about HERE.

PUT TABLE 1 about HERE.

There were 22,499 predictions made: true positives (1685), false positives (5996), false negatives (1556) and true negatives (13,262). The prevalence of the outcomes (repeat self-harm or suicide attempt) and the accuracy statistics for clinician prediction for each individual evaluation can be seen in Table 2 and LRs and CUI values seen in Table 3. The 12 month prevalence for general hospital populations ranged from 13% to 15% in the studies of individuals and was 17% in the evaluation using events. The prevalence in the psychiatric hospital populations was higher despite shorter follow-up, 21% (adult 3 month) and 17% (adolescent 1 month) and for the veterans' hospital lower despite longer follow-up, 7.1% (24 month).

PUT TABLE 2 ABOUT HERE.

PUT TABLE 3 ABOUT HERE.

PUT FIGURE 2a, 2b, 3a and 3b ABOUT HERE.

The meta-analysis yielded pooled estimates for sensitivity of 0.31 (95% CI: 0.18-0.50) and for specificity of 0.85 (0.75-0.92). We calculated a pooled PPV of 0.22 (0.21-0.23) and NPV of 0.89 (0.86-0.92), which can be seen in Figures 2a, 2b, 3a and 3c.

Secondary analyses

A secondary analysis using only those studies where individuals (not events) were followed over time (k=7) showed calculated estimates of PPV 0.22 (0.20-0.24) and NPV 0.89 (0.85-0.92).

Risk of bias

The QUADAS risk of bias ratings showed that only three of eight studies had a low risk of bias for patient selection, however there was a low risk of bias overall for the other three ratings (index test, reference standard and flow and timing). Details of risk of bias ratings for each study can be seen in Table 4.

PUT TABLE 4 ABOUT HERE.

PUT FIGURE 4a and 4b ABOUT HERE.

The estimates of between-study heterogeneity for sensitivity and specificity were large (I^2 = 98% and 100%, respectively). The between study heterogeneity for PPV was low (I^2 = 2%) but for NPV it was large (I^2 = 95%). As noted earlier, these values are somewhat misleading

in the context of meta-analyses evaluating the diagnostic or predictive properties of an instrument.

Discussion

Main findings

The question: "Can clinicians predict future self-harm with sufficient accuracy to be clinically valuable?" can be addressed by several results from our study. Firstly, sensitivity and specificity are useful accuracy statistics to understand how well a predictive test identifies cases and non-cases within a clinical population. The pooled estimates for sensitivity (0.31) shows that nearly 70% of all cases of future self-harm would be misclassified by clinicians as "low risk", meaning that most future cases would not be allocated to receive after-care aimed at reducing future self-harm, based on this risk classification. The pooled estimates for specificity (0.86) show that although 86% of future non-cases (i.e. no future self-harm event in the follow-up period) will be accurately classified as "low risk", 14% of the non-cases would therefore be misclassified as "high risk" (false positives) and allocated to after-care aimed at reducing future self-harm from which they would not benefit. These statistics are of more use at the public health level (e.g., to allocate resources needed for treatment at the hospital level) but are not able to accurately estimate the probability of future self-harm at the individual level, which is required by the clinician in order to make a binary judgement about allocation of after-care and "safe to discharge home" decisions.

There are several questions that can be used to guide decisions about the usefulness of a predictive test for the clinician. The first question would be: "Can a "high risk" classification be used to allocate after-care interventions for those that need it?" The pooled PPV (0.22) shows that only 22% of patients classified "high risk" will experience the adverse outcome of interest (i.e. self-harm). Thus, allocation of relatively expensive after-care (psychosocial interventions) with demonstrated effectiveness for reducing future self-harm (Hetrick, Robinson, Spittal, & Carter, 2016) based on this classification, will allocate 77% of the "high risk" group to after-care that will be of limited or no benefit for the future self-harm outcome. Another approach is to use the LR+ (1.67), which shows that there is very little change from the pre-test probability (14.8%) to the post-test probability (23%). This would be considered to be a minimal improvement in prediction over the expected rate and not clinically useful. A third approach is to use the CUI+ (0.11) (sensitivity x PPV), which would be considered to be

poor utility for a diagnostic or predictive test (Mitchell, 2011). These two additional tests of clinical utility reinforce the interpretation of the PPV result and indicate that a classification as "high risk" is not a useful basis for the allocation of psychosocial after-care interventions aimed at prevention of future self-harm.

Clinicians can also use the classification as "low risk" by a risk assessment to answer the important clinical question: "Is it safe to discharge this patient home?" For this question the clinician may reference the pooled NPV (0.89), which indicated that more than 1 in 10 patients classified as being "low risk" will have a future self-harm event, often soon after discharge (Carroll, Metcalfe, & Gunnell, 2014; Gunnell, Hawton, Ho, Evans, O'Connor et al, 2008). Hospital treated self-harm is a very common phenomena at the whole of population level and so this 10% misclassification as "low risk" means that many thousands of patients, at a national level, declared as "safe for discharge" will repeat self-harm, without the benefit of potentially effective after-care. Indeed, the LR- (0.70) indicates a small improvement from the pre-test probability (14.8%) to the post-test probability (11%). Similarly, the CUI- (0.62) (specificity x NPV) was considered to have only satisfactory utility because only 62% of non-cases would fall into a low risk categorization, despite this low risk judgement being a true negative in 89% of all low risk estimates. Although a nearly 90% accuracy in predicting future self-harm based on a negative test result might seem superficially reassuring, this is really only slightly different to the known probability of future self-harm in these populations before the clinician prediction is made (e.g. 86% with no repetition of hospital-treated self-harm after 12 months).

Confounding by Indication

There are an important set of questions to be considered when interpreting the PPV results (and if we ignore for this point the nearly 70% of repeat slef-harm cases that were allocated to the "low risk" classification).

Could the low PPV estimates be adversely affected because of confounding by indication (i.e. patients allocated to a high risk category then received after-care that was effective at reducing the rate of repetition in the high risk group)? If there was such confounding by indication, what might be the magnitude of such confounding? (i.e. by how much would interventions reduce the subsequent repeat self-harm rate and would this substantially affect the interpretation of the outcomes of the study)?

Firstly, all of the original studies seem to infer that risk assessment was part of usual clinical practice and so we must assume that the possibility of confounding by indication is real. Secondly, the allocation of after-care and the compliance of hospital treated self-harm

patients in attending at allocated after-care interventions is generally poor. In a study of all Norwegian hospitals, which have a commitment to a codified high standard of care, 76% of hospitals met the standard for allocation of after-care (to 90% or more of all suicide attempt patients) (Mehlum, Mork, Reinholdt, Fadum, & Rossow, 2010). Historically, Kreitman reported that 50% of hospital-treated self-harm patients failed to attend any outpatient appointment (Kreitman, 1979), whilst O'Brien reported 60% non-compliance at one week and 69% non-attendance after four weeks for allocated outpatient treatment (O'Brien, Holton, Hurren, Watt, & Hassanyeh, 1987). Even in more recent times and with assertive outreach to a highly selected sample, for those offered after-care only 73% attended one or more sessions (Murphy, Steeg, Cooper, Chang, Turpin et al, 2010). Thirdly, we need to consider the potential effect size for those allocated to and compliant with an allocated after-care intervention. A recent systematic review estimated that the absolute risk reduction for any repeat self-harm for all psycho-social interventions combined was 2.5% (18.0% v 20.5%) Even when this estimate was restricted to the most effective intervention (five CBT studies) the absolute risk reduction was 9.9% (31.2% v 41.1%). (Hetrick, Robinson, Spittal, & Carter, 2016). Even a 10 % absolute reduction in repetition for the 7.6%-19.9% of patients allocated to the "high risk" group in our studies, would only increase the overall prevalence in repetition of self-harm by 1-2%. We do not know of any randomised controlled trial that could be used to assess the accuracy of these estimates. However a recent non-randomised multi-centre trial from eight EDs in the US compared a universal screening plus intervention (secondary suicide risk screening by the ED physician, discharge resources, and post-ED telephone calls) to treatment as usual, with one year follow-up, a combined endpoint of any non-fatal suicide attempt or suicide death, with a reported prevalence rate for this endpoint of 20.9% (Miller, Camargo, Jr, & Arias, 2017). Patients in the intervention phase showed a 5% absolute reduction in suicide attempt (23% vs 18%), and about 8% (49% vs 41%) for a wider combined suicidal behaviour endpoint; which was less than the pooled estimate of 10% absolute reduction for the five CBT interventions. So confounding by indication, even with full treatment compliance, and the highest proven treatment response, would not improve the PPV sufficiently to be clinically useful.

Ultimately however, it is important to understand that low prevalence outcomes impose a statistical ceiling on PPV (and a statistical limitation of LR+ values), which is independent of how the risk classification was made (Carter, Milner, McGill, Pirkis, Kapur et al, 2017).

Prediction and prevalence

The use of clinician prediction for future clinical outcomes can be useful in some situations, although this is highly dependent on the prevalence (or pre-test probability) of the outcome being predicted. Low prevalence disorders impose a statistical ceiling on PPV and NPV, which means that prediction will not be accurate enough to be the basis on which to allocate or not allocate treatment to an individual patient. The ideal situation is for a pre-test probability of 50%, which means that statistically PPV and NPV might be clinically useful if the prediction (post-test probability) is accurate enough. A recent successful example of this approach has been in the clinician prediction of time to death in donor-eligible ICU patients (Brieva, Coleman, Lacey, Harrigan, Lewin et al, 2014).

However, prediction of future events like suicide cannot be adequately predicted because the event is of low prevalence (Rosen, 1954); and in this study we have shown that clinician prediction of future self-harm, which has a higher prevalence than suicide but is still a low prevalence behaviour, is also not clinically useful for the allocation of after-care. In essence, clinician risk classification for future self-harm offers no meaningful improvement to the pre-test probability for any individual patient classified as "high risk".

In order to achieve what we would consider to be the *minimal* threshold for a clinically useful level of accuracy based on likelihood ratios of LR+ 5.00 and LR- 0.1, we have calculated that for a pre-test probability of 14.4% and *n*=22,499 patients, classification of high versus low risk would have the following properties: sensitivity 0.92, specificity 0.82, PPV 0.46, NPV 0.98, which would still result in the inappropriate allocation of after-care intervention to 3,537 patients who would not be at risk of future self-harm and so could not benefit from that aspect of intervention; whilst failing to allocate after-care to 264 patients who would have future self-harm. Should this hypothetical level of test accuracy ever be achieved, non-allocation of after-care could be done accurately for 15,722 patients, and this would represent a substantial reduction in after-care costs compared to a system of allocation of after-care to all patients; with the additional "harm" of 264 (8%) future cases being misclassified as "low risk" and so not offered potentially beneficial after-care. However, we would conclude that the usual prevalence of future self-harm (16%) in hospital-treated slef-harm populations and (6%) in psychiatric inpatient populations, is simply too low to allow for clinically useful prediction in these two important hospital-treated populations.

Changing current practice

The NICE Clinical Guideline # 133 in the UK recommends that risk assessment tools and scales to predict future suicide or repetition of self-harm should not be used as the basis to allocate after-care (National Institute for Health and Care Excellence, 2011), as does the

Royal Australian and New Zealand College of Psychiatrists (Carter, Page, Large, Hetrick, Milner et al, 2016). Instead, recommendations are made for a needs based approach combined with an assessment of modifiable risk factors to inform possible allocation of after-care on an individual patient basis (subject to agreement from the patient). We would endorse this approach for the allocation of after-care interventions. An excellent summary of the evidence base and potentially useful interventions has been proposed by the Center for Disease Control in the US (Stone, Holland, Bartholow, Crosby, Davis et al, 2017). After-care interventions for those discharged home should be aimed at meeting patient needs and by working with relevant community-based services (e.g. safety plans, personal support, resolution of interpersonal difficulties, social function, occupational function, vocational support, housing and financial needs); reducing exposure to modifiable clinical risk factors for self-harm and repeat self-harm (e.g. mental illness, substance misuse, personality disorder); and specific psychosocial therapies for reducing repetition of self-harm (Hetrick, Robinson, Spittal, & Carter, 2016; National Health and Medical Research Council., 2012). In a review of UK general hospitals, guidelines for the assessment of needs were available for patients who self-harm in 7 of the 32 (21.7%) Emergency Departments and 29 of the 32 (90.6%) mental health services as part of the patients' wider assessment process (Quinlivan, Cooper, Steeg, Davies, Hawton et al, 2014). If this is a representative estimate of needs assessment at the national level in the UK, there is still considerable scope for improvement using this approach as the basis for allocation of after-care interventions, at least in the general hospital.

Little is known about how clinicians decide a patient is not "safe for discharge". The decision to transfer to a psychiatric hospital for continuing care after an episode of hospital-treated self-harm is associated with many factors; with higher levels of current suicidal ideation and suicidal planning being the strongest independently associated variables (Carter, Safranko, Lewin, Whyte, & Bryant, 2006). However, the "appropriateness" of this decision to transfer to and then admit to a psychiatric hospital cannot be determined; and the effectiveness of psychiatric hospitalisation to reduce future suicide and self-harm has never been tested in a randomised controlled trial. It is likely that changing the reliance by clinical services on risk classification to allocate after-care and to make the "safe to discharge" decisions will take some time, whether the outcome of interest is subsequent suicide or self-harm.

Strengths and Limitations

This study was based on a systematic search and review using the PRISMA standard. Quality ratings and risk of bias were conducted according to the QUADAS-2 standards, which is the appropriate instrument for studies of diagnostic precision or predicting future outcomes. Meta-analysis using appropriate models was conducted to produce pooled estimates of sensitivity, specificity, PPV and NPV and to estimate between-study heterogeneity and potential publication bias. Clinical usefulness was comprehensively assessed by considering the sensitivity, specificity, PPV, NPV, LR+, LR-, CUI+ and CUI- values for individual evaluations. This is the first meta-analysis study of the accuracy of clinician prediction of future suicidal behaviours.

It is possible that our search missed some studies and it is possible that there may be a non-publication bias for studies that failed to produce adequate prediction. The risk of bias in the selected evaluations was modest and the number of evaluations was also modest (8 evaluations on 22,499 occasions) and so the specific estimates must be viewed with a degree of caution. Future large-scale studies could potentially change the reported pooled estimates. The prevalence rates of future self-harm in the selected evaluations were similar, however caution should be used in applying these result in populations with markedly different prevalence rates. Populations with lower prevalence rates will have even more dismal accuracy.

Conclusions

Simply stated, for clinician classification of "high risk" for repeated self-harm, the pooled sensitivity (0.31) indicates about 70% of repeater cases will be missed and the pooled PPV (.22) indicates that a high risk classification for repetition of self-harm will be incorrect nearly 80% of the time. This level of inaccuracy is just too high for this approach to be clinically useful to allocate or differentially allocate interventions to reduce the predicted outcome of repetition of self-harm. Clinician prediction of future self-harm is not adequately accurate to determine allocation of after-care or safety to discharge. Given the low prevalence of future self-harm as an outcome, even in clinically high risk populations, prediction by any method will not be clinically useful at the individual patient level. Instead we would recommend three approaches to allocation of clinical after-care: an individual needs based assessment aimed at reducing exposure to modifiable risk factors (Ryan & Large, 2013; Carter, Page, Large, Hetrick, Milner et al, 2016); allocation of proven interventions for particular selected high-risk sub-populations e.g. Borderline Personality Disorder (National Health and Medical Research Council., 2012); and allocation of proven interventions that can

be delivered to unselected high risk clinical populations (Hetrick, Robinson, Spittal, & Carter, 2016; Milner, Carter, Pirkis, Robinson, & Spittal, 2015). There is currently no sufficiently accurate way to determine which patient is “safe for discharge”, whether the outcome of interest is suicide or self-harm.

Acknowledgements

Acknowledgements

Katherine McGill’s position is funded by the Burdekin Suicide Prevention Program and administered by Hunter New England Mental Health Services.

Declaration of Interests

Navneet Kapur chaired the NICE guidelines for the longer term management of self-harm in England but the views in this paper are the author’s own and not those of NICE or the Department of Health (UK). Gregory Carter chaired the Royal Australian and New Zealand College of Psychiatrists Clinical Practice Guidelines for Deliberate Self Harm but the views in this paper are the author’s own and not those of the RANZCP.

Table 1: Characteristics of evaluations of clinician prediction studies

Author	Year	Setting[a]	Country	Clinician[b]	Gender	Outcome[c]	Method	FU	Intake	Retain	High v Low	High (%)	High/Med/Low
Erdman	1987	PIS	USA	MH	M	SA	Clinician	3 mth	61	52	7 v 45	(13.5)	NA
Kapur	2005	ED	UK	ED	M/F	SH	Records	12 mth	4879	4879	971 v 3908	(19.9)	971/2284/1624
Kapur	2005	ED	UK	MH	M/F	SH	Records	12 mth	3828	3828	464 v 3913	(10.6)	369/1738/1721
Cooper	2007	ED	UK	MH	M/F	SH/SD	Records	6 mth	*9086	*8722	5736 v 2986	(65.8)	NA[d]
Murphy	2011	ED	UK	MH Nurse	M/F	SH	Records	12 mth	2626	2626	227 v 1860	(10.9)	NA
Murphy	2011	ED	UK	Trainee Psychiatrist	M/F	SH	Records	12 mth	865	865	48 v 581	(7.6)	NA
Phillips	2012	PIS	Australia	Psychiatrist	M/F	SH	Records	1 mth	193	156	32 v 161	(16.6)	32/124/37[e]
Naydock	2015	ED (VA)	USA	MH	M/F	SA	Records	24 mth	1560	1560	196 v 1364	(12.6)	NA[f]

[a] PIS= Psychiatric Inpatient service; ED= Emergency Department (general hospital); ED (VA)= Emergency Department (veterans administration hospital)

[b] MH= Mental Health Professional; ED= Emergency Department staff

[c] SA: suicide attempt. SH: self-harm, SD: suicide death

[d] Authors combined 'medium' with 'high' risk

[e] Authors originally used a 'high' risk and 'very high' risk stratification

[f] Authors combined 'moderate' with 'low' risk

* unit of analysis is events not individuals

Table 2: Prevalence of repeat self-harm and accuracy statistics for clinician classification for future self-harm

Author	Year	Prevalence n (%)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
Erdman	1987	11 (21.2)	0.36 (0.11-0.69)	0.93 (0.80-0.98)	0.57 (0.18-0.90)	0.84 (0.71-0.94)
Kapur	2005	646 (13.2)	0.32 (0.28-0.36)	0.82 (0.81-0.83)	0.21 (0.19-0.24)	0.89 (0.88-0.90)
Kapur	2005	549 (13.0)	0.17 (0.14-0.21)	0.90 (0.89-0.91)	0.20 (0.17-0.24)	0.88 (0.87-0.89)
Cooper	2007	* 1481 (17.0)	0.85 (0.83-0.87)	0.38 (0.37-0.39)	0.22 (0.21-0.23)	0.92 (0.91-0.93)
Murphy	2011	320 (15.3)	0.17 (0.13-0.22)	0.90 (0.89-0.92)	0.25 (0.19-0.31)	0.86 (0.84-0.87)
Murphy	2011	93 (14.8)	0.12 (0.06-0.20)	0.93 (0.91-0.95)	0.23 (0.12-0.37)	0.86 (0.83-0.89)
Phillips	2012	31 (16.6)	0.26 (0.12-0.45)	0.85 (0.79-0.90)	0.25 (0.11-0.43)	0.86 (0.79-0.91)
Naydock	2015	110 (7.1)	0.45 (0.35-0.54)	0.90 (0.88-0.91)	0.25 (0.19-0.32)	0.96 (0.94-0.97)

* unit of analysis is events not individuals

PPV = Positive Predictive Value

NPV = Negative Predictive Value

(95% CI) = Confidence Interval 95%

Table 3: Other measures of clinical utility for clinician classification for future self-harm

Author	Year	Prevalence n (%)	LR+ (95%CI)	LR- (95%CI)	CUI+ (95%CI)	CUI- (95%CI)
Erdman	1987	11 (21.2)	4.97 (1.30-19.00)	0.69 (0.44-1.08)	0.21 (0.00-0.61)	0.78 (0.70-0.86)
Kapur	2005	646 (13.2)	1.78 (1.56-2.02)	0.83 (0.79-0.88)	0.07 (0.03-0.11)	0.73 (0.72-0.74)
Kapur	2005	549 (13.0)	1.80 (1.46-2.21)	0.92 (0.88-0.95)	0.04 (0.00-0.08)	0.80 (0.79-0.81)
Cooper	2007	* 1481 (17.0)	1.37 (1.33-1.41)	0.40 (0.35-0.45)	0.19 (0.17-0.21)	0.35 (0.34-0.36)
Murphy	2011	320 (15.3)	1.81 (1.37-2.39)	0.91 (0.87-0.96)	0.04 (0.00-0.10)	0.78 (0.76-0.79)
Murphy	2011	93 (14.8)	1.71 (0.91-3.24)	0.95 (0.88-1.02)	0.03 (0.00-0.13)	0.80 (0.78-0.82)
Phillips	2012	31 (16.1)	1.74 (0.86-3.51)	0.87 (0.70-1.08)	0.06 (0.00-0.25)	0.73 (0.68-0.78)
Naydock	2015	110 (7.1)	4.39 (3.39-5.69)	0.62 (0.52-0.73)	0.11 (0.02-0.20)	0.86 (0.85-0.87)

Pooled (unweighted)	3241 (14.4)	1.67 (1.61-1.74)	0.70 (0.67-0.72)	0.11 (0.10-0.13)	0.62 (0.61-0.62)

* unit of analysis is events not individuals

LR+ Likelihood Ratio for a positive test

LR- Likelihood Ratio for a negative test

CUI+ Clinical Utility Index for a positive test

CUI- Clinical Utility Index for a negative test

(95% CI) = Confidence Interval 95%

Table 4: QUADAS ratings of risk of bias

Study	Year	RISK OF BIAS			
		PATIENT SELECTION	INDEX TEST	REFERENCE STANDARD	FLOW AND TIMING
Erdman	1987	☹	?	?	☹
Kapur	2005[a]	☺	☺	☺	☺
Kapur	2005[b]	☺	☺	☺	☺
Cooper	2007	☹	☺	☺	☺
Murphy	2011[a]	☹	☺	☺	☺
Murphy	2011[b]	☹	☺	☺	☺
Phillips	2012	☺	☺	?	☺

Naydock	2015	☹	☺	?	☹

☺Low Risk ☹High Risk ? Unclear Risk

Figure 1: PRISMA diagram

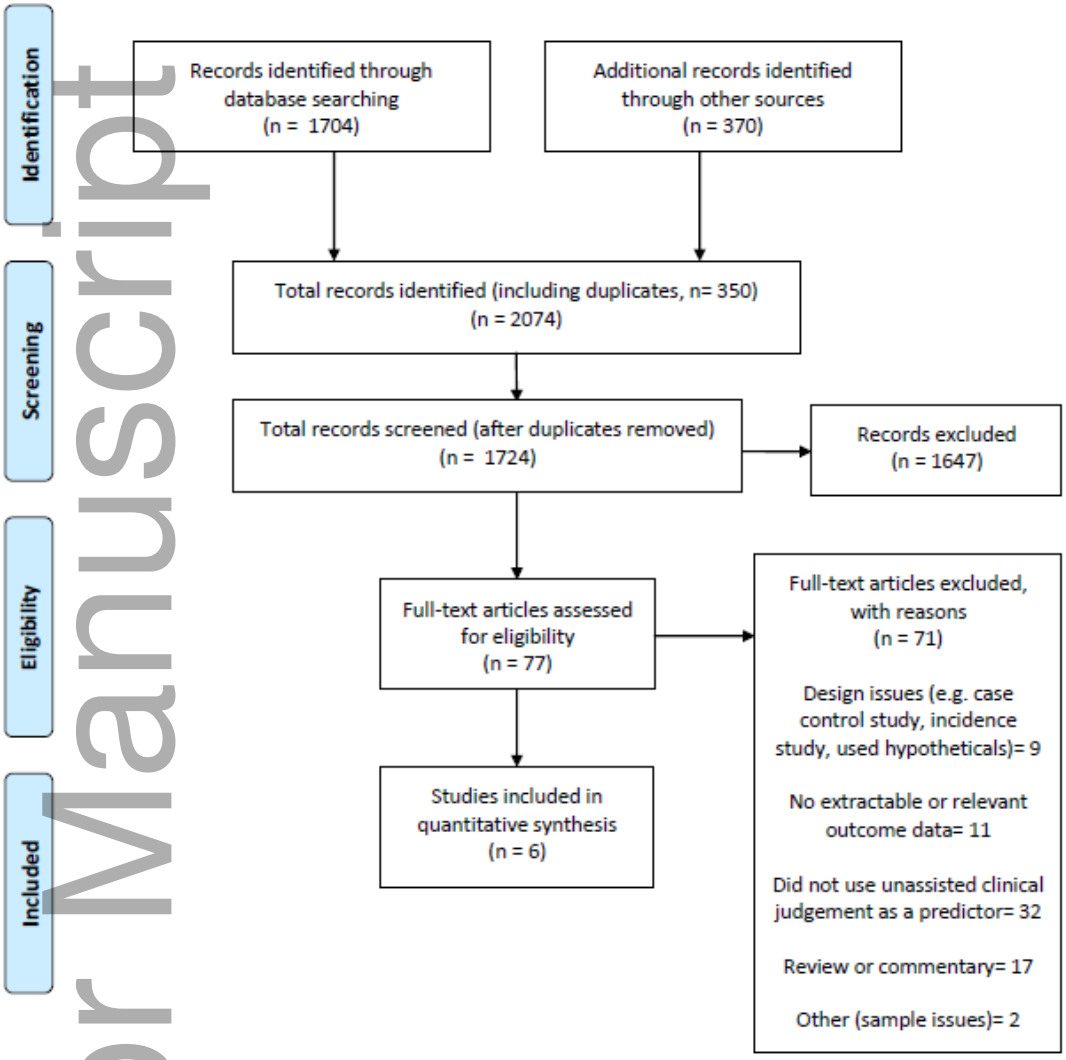


Figure 2a: Forest plots for pooled estimates of Sensitivity

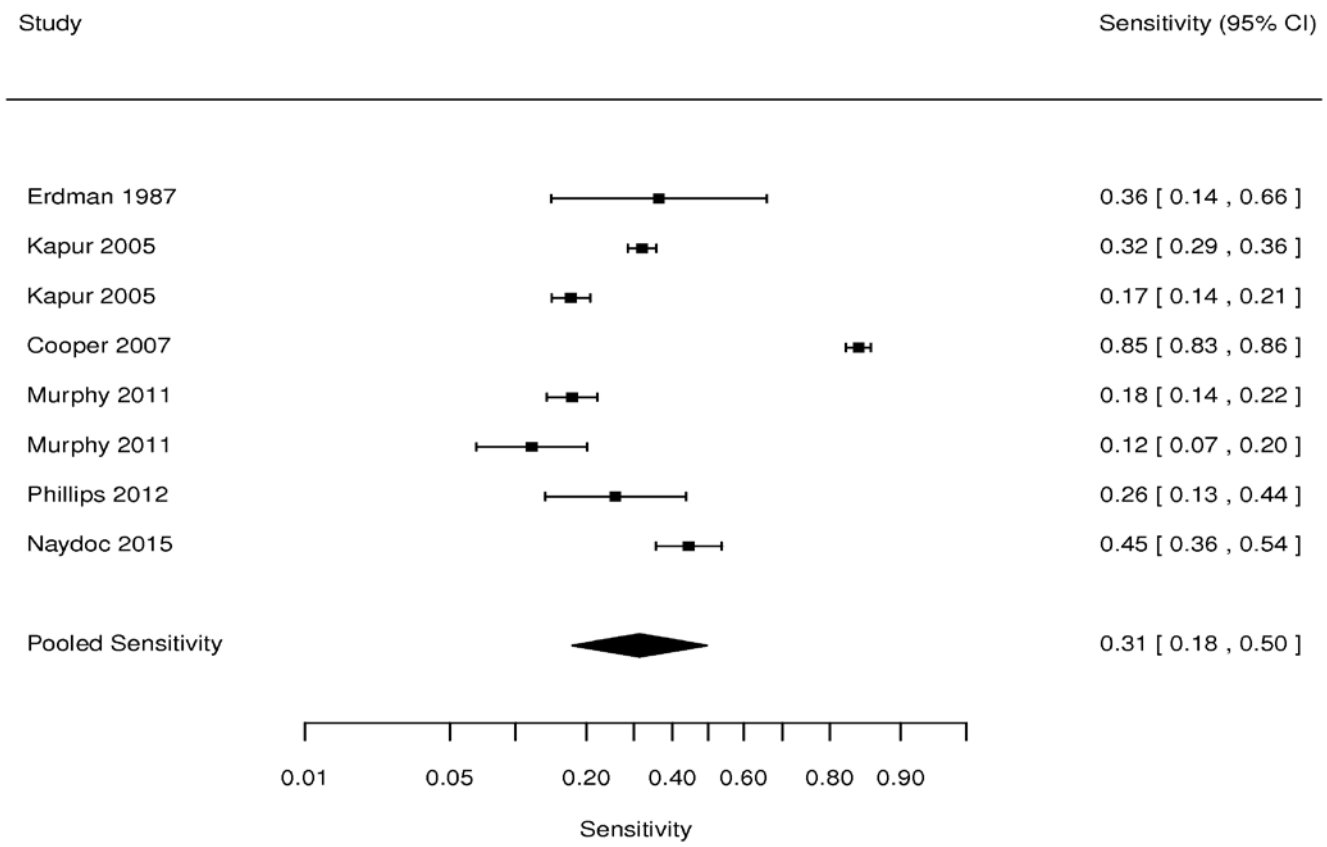


Figure 2b: Forest plots for pooled estimates of Specificity

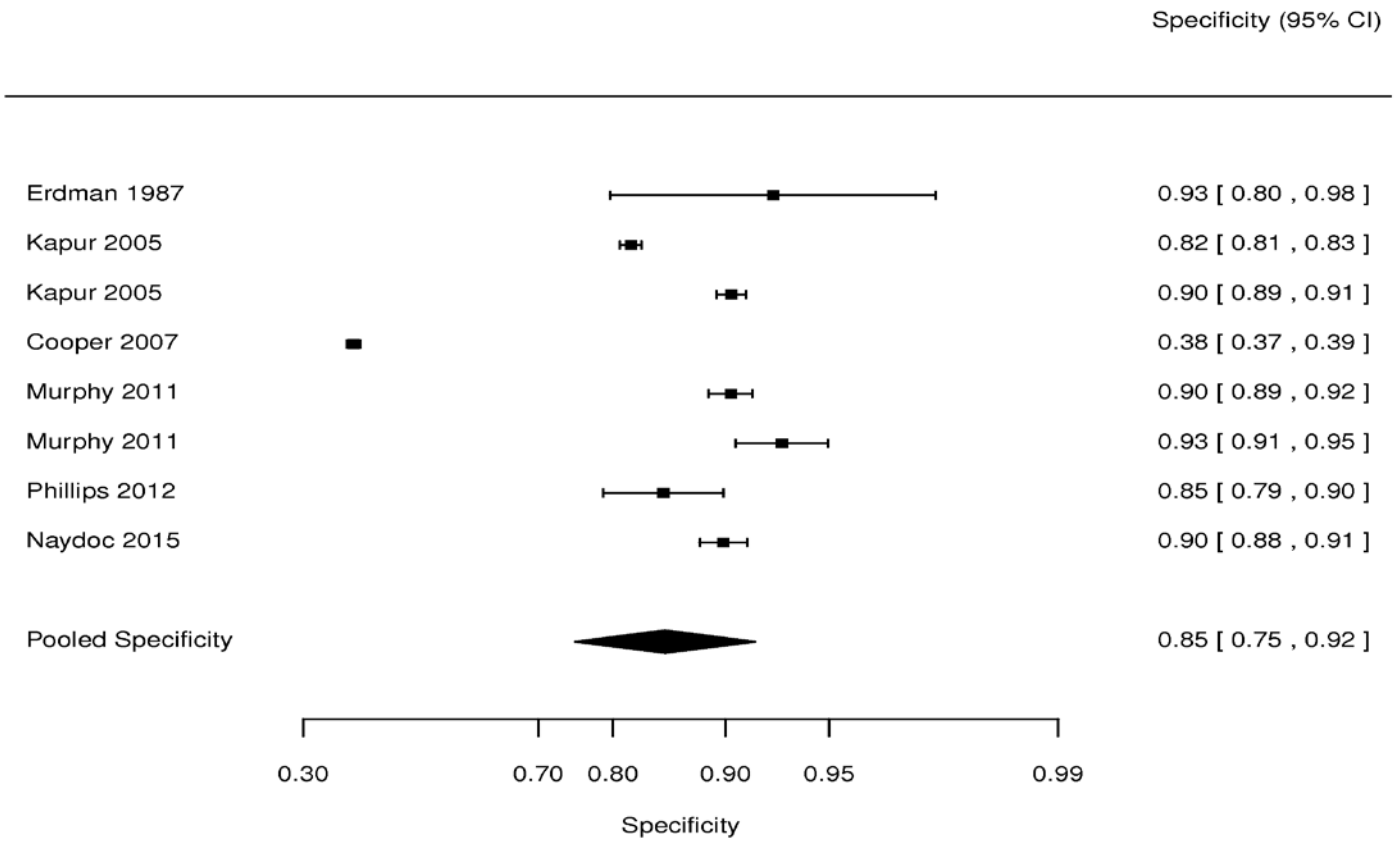


Figure 3a: Forest plots for estimates for Positive Predictive Values

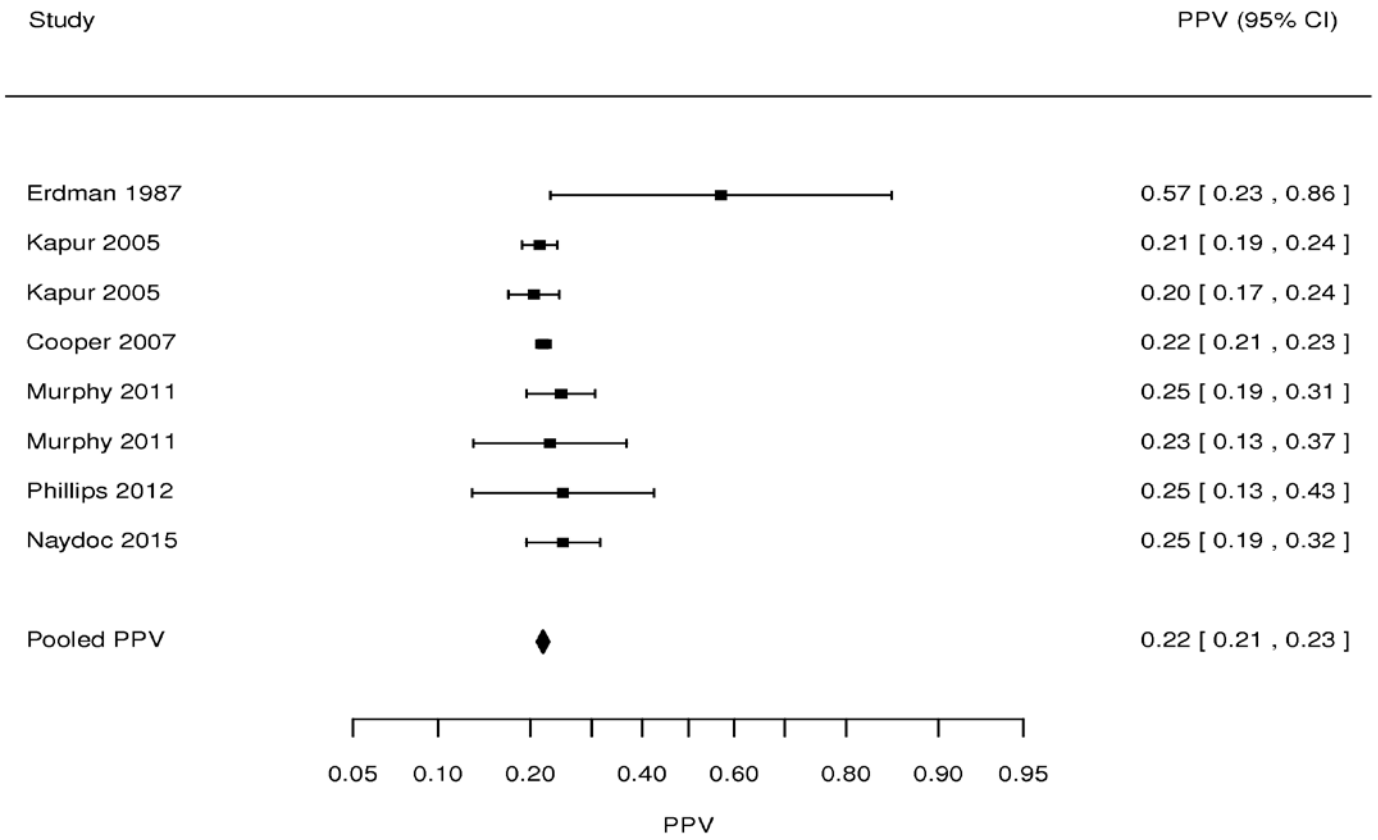


Figure 3b: Forest plots for estimates for Negative Predictive Values

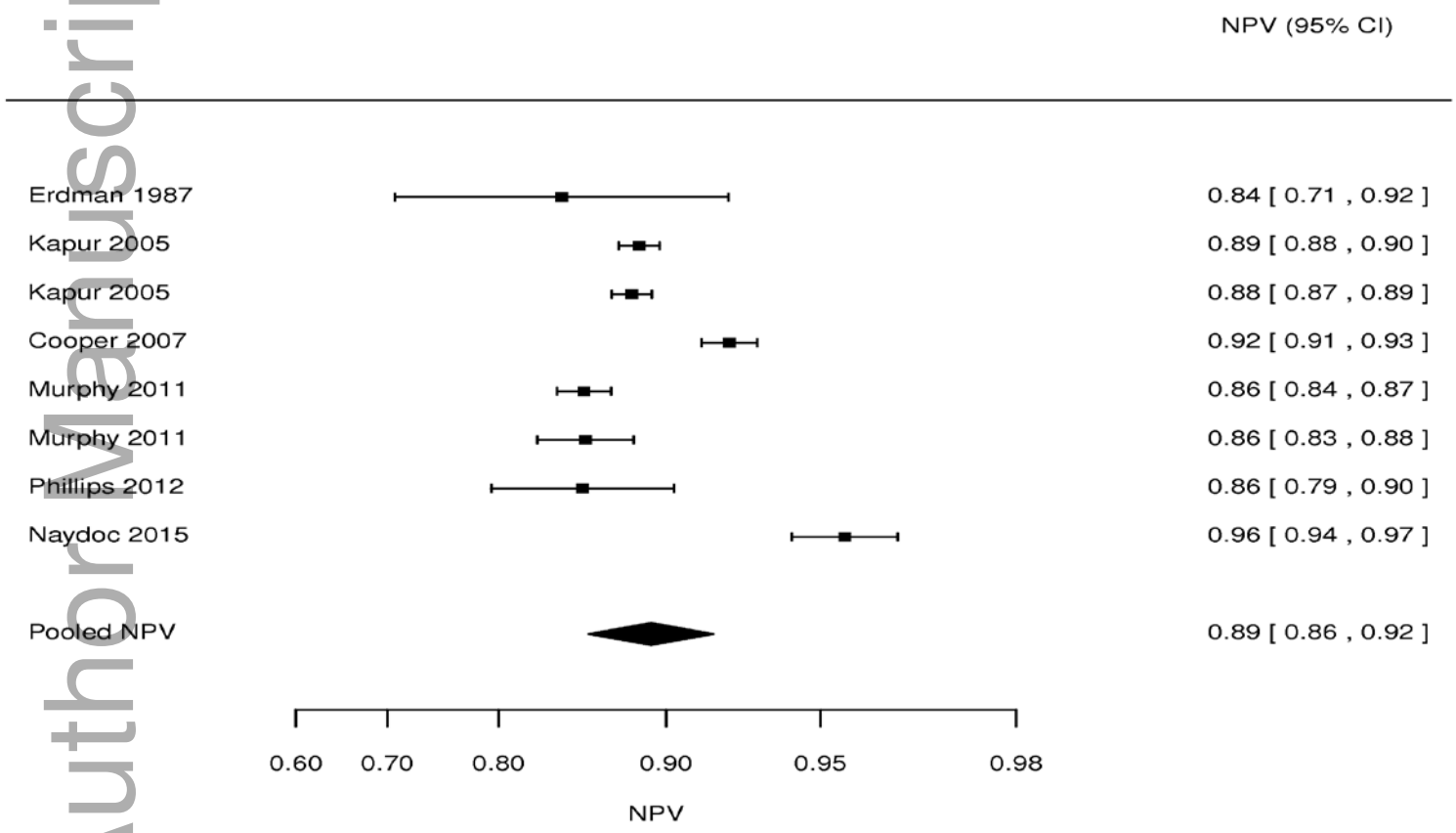


Figure 4a: Funnel Plots for Sensitivity and Specificity

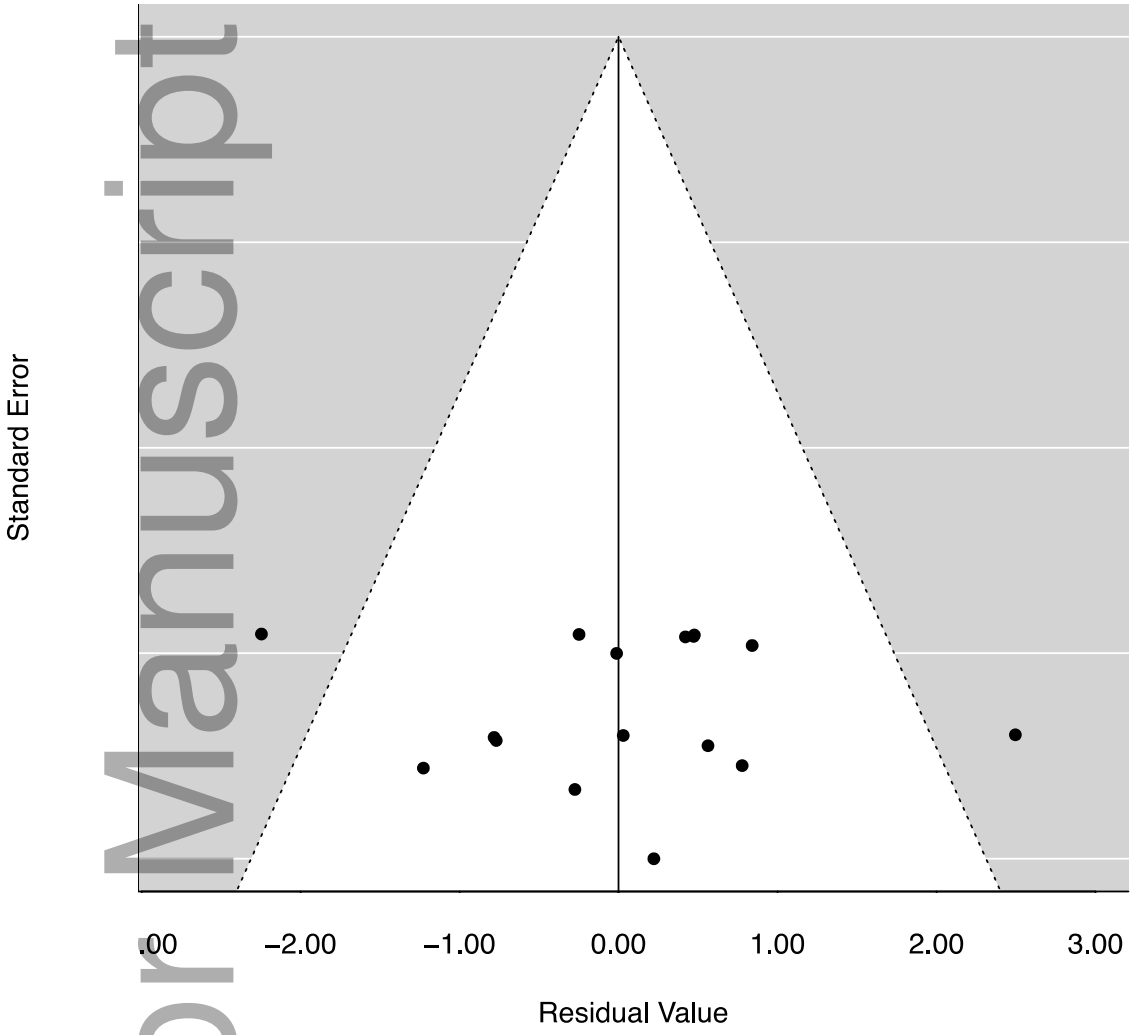
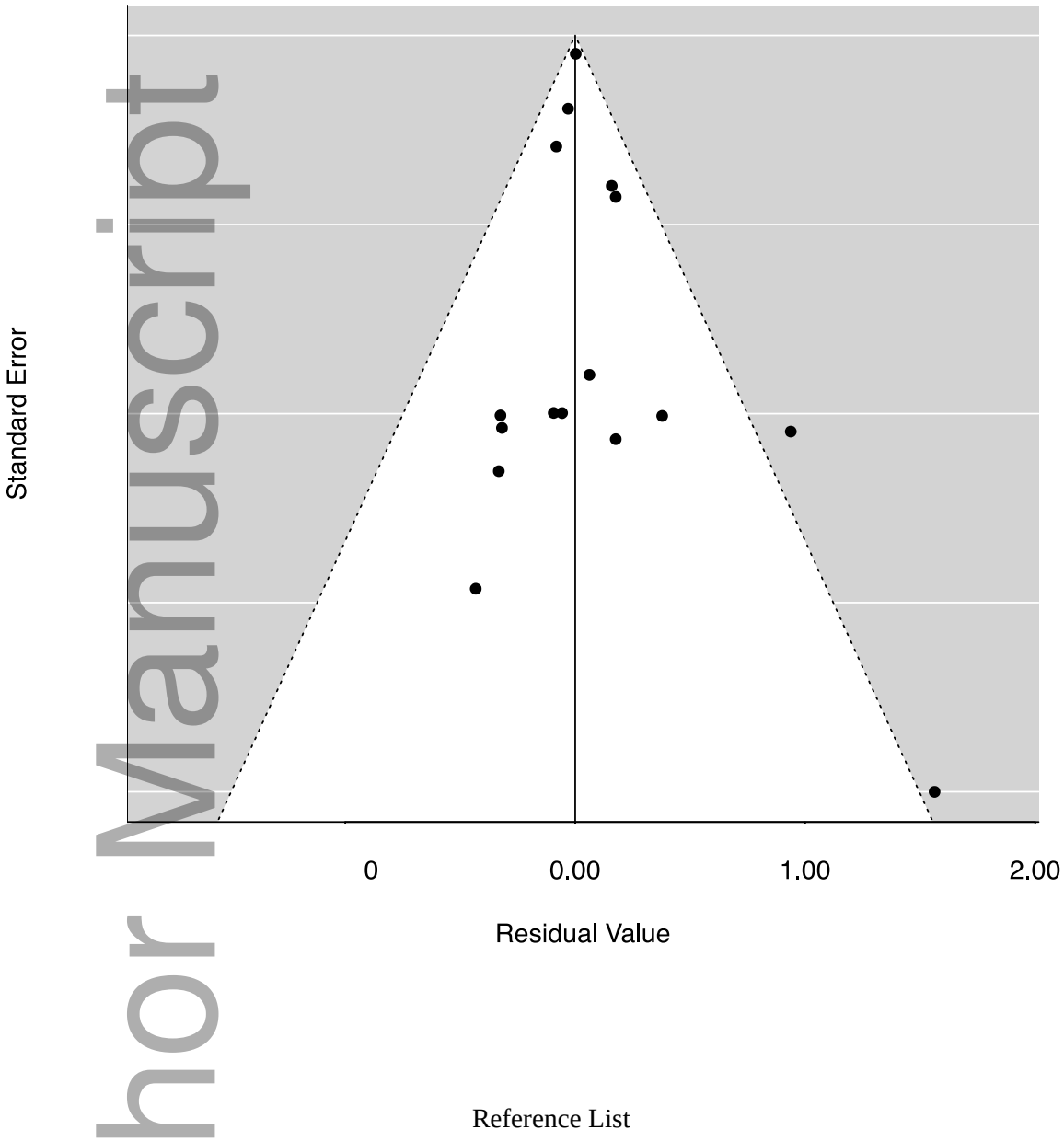


Figure 4a: Funnel Plots for PPV and NPV



Reference List

BOSSUYT, P., DAVENPORT, C., DEEKS, J., HYDE, C., LEEFLANG, M., & SCHOLTEN, R. (2013). Interpreting results and drawing conclusions Cahpter 11. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 0.9*

BRIEVA, J., COLEMAN, N., LACEY, J., HARRIGAN, P., LEWIN, T. J., & CARTER, G. L. (2014). Prediction of Death in Less Than 60 Minutes After Withdrawal of Cardiorespiratory Support in Potential Organ Donors After Circulatory Death. *Transplantation, 98*, 1112-1118.

CARROLL, R., METCALFE, C., & GUNNELL, D. (2014). Hospital Presenting Self-Harm and Risk of Fatal and Non-Fatal Repetition: Systematic Review and Meta-Analysis. *PLoS ONE, 9*, e89944

CARTER, G., MILNER, A. J., MCGILL, K., PIRKIS, J., KAPUR, N., & SPITTAL, M. (2017). Predicting suicidal behaviours using clinical instruments: systematic review and meta-analysis of positive predictive values for risk scales. *British Journal of Psychiatry, 6*, 387-395.

CARTER, G. L., SAFRANKO, I., LEWIN, T. J., WHYTE, I. M., & BRYANT, J. L. (2006). Psychiatric hospitalization after deliberate self-poisoning. *Suicide and Life-Threatening Behavior, 36*, 213-222.

CARTER, G., PAGE, A., LARGE, M., HETRICK, S., MILNER, A. J., BENDIT, N. et al. (2016). Royal Australian and New Zealand College of Psychiatrists clinical practice guideline for the management of deliberate self-harm. *Australian and New Zealand Journal of Psychiatry, 50*, 939-1000.

CHAN, M. K. Y., BHATTI, H., MEADER, N., STOCKTON, S., EVANS, J., CONNOR, R. C. et al. (2016). Predicting suicide following self-harm: systematic review of risk factors and risk scales. *The British Journal of Psychiatry, 209*, 277-283.

CHAPMAN, C. L., MULLIN, K., RYAN, C. J., KUFFEL, A., NIELSSEN, O., & LARGE, M. M. (2015). Meta-analysis of the association between suicidal ideation and later suicide among patients with either a schizophrenia spectrum psychosis or a mood disorder. *Acta Psychiatrica Scandinavica, 131*, 162-173.

COOPER, J., KAPUR, N., & MACKWAY-JONES, K. (2007). A comparison between clinicians' assessment and the Manchester Self-Harm Rule: a cohort study. *Emergency Medicine Journal, 24*, 720-721.

EGGER, M., SMITH, G. D., SCHNEIDER, M., & MINDER, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ, 315*, 629

ERDMAN, H. P., GREIST, J. H., GUSTAFSON, D. H., TAVES, J. E., & KLEIN, M. H. (1987). Suicide Risk Prediction by Computer Interview - A Prospective-Study. *Journal of Clinical Psychiatry, 48*, 464-467.

FRANKLIN, J. C., RIBEIRO, J. D., FOX, K. R., BENTLEY, K. H., KLEIMAN, E. M., HUANG, X. et al. (2017). Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological Bulletin, 143*, 187-232.

GUNNELL, D., HAWTON, K., HO, D., EVANS, J., O'CONNOR, S., POTOKAR, J. et al. (2008). Hospital admissions for self harm after discharge from psychiatric inpatient care: cohort study. *BMJ, 337*,

HARBORD, R. M., EGGER, M., & STERNE, J. A. (2006). A modied test for small-study eects in meta-analyses of controlled trials with binary endpoints. *Statistics in Medicine*, 3443-3457.

HETRICK, S. E., ROBINSON, J., SPITTAL, M. J., & CARTER, G. (2016). Effective psychological and psychosocial approaches to reduce repetition of self-harm: a systematic review, meta-analysis and meta-regression. *BMJ Open, 6*,

HUBERS, A. A. M., MOADDINE, S., PEERSMANN, S. H. M., STIJNEN, T., VAN DUIJN, E., VAN DER MAST, R. C. et al. (2016). Suicidal ideation and subsequent completed suicide in both psychiatric and non-psychiatric populations: a meta-analysis. *Epidemiology and Psychiatric Sciences, 2016/12/19*, 1-13.

KAPUR, N., COOPER, J., RODWAY, C., KELLY, J., GUTHRIE, E., & MACKWAY-JONES, K. (2005). Predicting the risk of repetition after self harm: cohort study. *BMJ, 330*, 394-395.

KREITMAN, N. (1979). Reflections on the management of parasuicide. *The British Journal of Psychiatry, 135*, 275

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

LIBERATI, A., ALTMAN, D. G., TETZLAFF, J., MULROW, C., GOTZSCHE, P. C., IOANNIDIS, J. P. A. et al. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *Bmj-British Medical Journal, 339*,

MACASKILL, P., GATSONIS, C., DEEK, J. J., HARBORD, R. M., & TAKWOINGI, Y. (2010). Analysing and Presenting Results Chapter 10. In J.J. Deeks, P.M. Bossuyt, & C. Gatsonis (Eds.), *Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0*.

MEHLUM, L., MORK, E., REINHOLDT, N., FADUM, E., & ROSSOW, I. (2010). Quality of Psychosocial Care of Suicide Attempters at General Hospitals in Norway–A Longitudinal Nationwide Study. *Archives of Suicide Research, 14,* 146-157.

MILLER, I. W., CAMARGO, C. A., JR, & ARIAS, S. A. (2017). Suicide prevention in an emergency department population: The ed-safe study. *JAMA Psychiatry, 74,* 563-570.

MILNER, A. J., CARTER, G., PIRKIS, J., ROBINSON, J., & SPITTAL, M. J. (2015). Letters, green cards, telephone calls and postcards: systematic and meta-analytic review of brief contact interventions for reducing self-harm, suicide attempts and suicide. *The British Journal of Psychiatry, 206,* 184

MITCHELL, A. J. (2011). Sensitivity x PPV is a recognized test called the clinical utility index (CUI+). *European Journal of Epidemiology, 26,* 251-252.

MURPHY, E., KAPUR, N., WEBB, R., & COOPER, J. (2011). Risk assessment following self-harm: comparison of mental health nurses and psychiatrists. *Journal of Advanced Nursing, 67,* 127-139.

MURPHY, E., STEEG, S., COOPER, J., CHANG, R., TURPIN, C., GUTHRIE, E. et al. (2010). Assessment Rates and Compliance With Assertive Follow-up After Self-Harm: Cohort Study. *Archives of Suicide Research, 14,* 120-134.

NATIONAL HEALTH AND MEDICAL RESEARCH COUNCIL. (2012). Clinical Practice Guideline for the Management of Borderline Personality Disorder.

NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE. (2011). NICE guidelines Self-harm: longer-term management [CG 133].

NAYDOCK, G. R. (2015). Naturalistic Clinical Decision Making by Emergency Department Staff and the Assignment of Categorical Suicide Risk ratings within an Urban Veteran Population.

O'BRIEN, G., HOLTON, A. R., HURREN, K., WATT, L., & HASSANYEH, F. (1987). Deliberate self-harm and predictors of out-patient attendance. *The British Journal of Psychiatry, 150,* 246

OWENS, D., HORROCKS, J., & HOUSE, A. (2002). Fatal and non-fatal repetition of self-harm. Systematic review. *British Journal of Psychiatry, 181,* 193-199.

PHILLIPS, N. L., STARGATT, R., & BROWN, A. (2012). Risk assessment of self- and other-directed aggression in adolescent psychiatric inpatient units. *Australian and New Zealand Journal of Psychiatry, 46,* 40-46.

QUINLIVAN, L., COOPER, J., STEEG, S., DAVIES, L., HAWTON, K., GUNNELL, D. et al. (2014). Scales for predicting risk following self-harm: an observational study in 32 hospitals in England. *BMJ Open, 4,*

RIBEIRO, J. D., FRANKLIN, J. C., FOX, K. R., BENTLEY, K. H., KLEIMAN, E. M., CHANG, B. P. et al. (2016). Self-injurious thoughts and behaviors as risk factors for future suicide ideation, attempts, and death: a meta-analysis of longitudinal studies. *Psychological Medicine, 2015/09/15,* 225-236.

ROSEN, A. (1954). Detection of Suicidal Patients: An Example of Some Limitations in the Prediction of Infrequent Events. *Journal of Consulting Psychology, 18,* 397-403.

ROYAL COLLEGE OF PSYCHIATRISTS. (2010). Self-harm, suicide, and risk: helping people who self-harm. Final report of a working group. *CR 158,*

RYAN, C. J. & LARGE, M. M. (2013). Suicide risk assessment: where are we now? *Medical Journal of Australia., 198,* 462-463.

SILVERMAN, M. M. & BERMAN, A. L. (2014). Suicide Risk Assessment and Risk Formulation Part I: A Focus on Suicide Ideation in Assessing Suicide Risk. *Suicide and Life-Threatening Behavior, 44,* 420-431.

STIJNEN, T., HAMZA, T. H., & OZDEMIR, P. (2010). Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Statistics in Medicine, 29,* 3046-3067.

STONE, D. M., HOLLAND, K. M., BARTHOLOW, B., CROSBY, A. E., DAVIS, S., & WILKINS, N. (2017). Preventing Suicide: A Technical Package of Policies, Programs, and Practices.

SUICIDE PREVENTION RESOURCE CENTER. (2015). Caring for Adult Patients with Suicide Risk: A Consensus Guide for Emergency Departments.

WHITING, P. F., RUTJES, A. W., WESTWOOD, M. E., MALLETT, S., DEEKS, J. J., REITSMA, J. B. et al. (2011). QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine, 155,* 529-536.

WOODFORD, R., CARTER, G., MCGILL, K., MILNER, A., PIRKIS, J., & SPITTAL, M. (2015). Accuracy of clinician predicted suicidal behaviour: a systematic review and meta-analysis. *PROSPERO, CRD42015015781*,