Use of a two-phase process to identify possible cases of mental ill health in the UK military

HOWARD BURDETT,¹ NICOLA T. FEAR,² NORMAN JONES,² NEIL GREENBERG,² SIMON WESSELY¹ & ROBERTO J. RONA¹

1 King's Centre for Military Health Research, Department of Psychological Medicine, King's College London, London, UK

2 Academic Department of Military Mental Health, Department of Psychological Medicine, King's College London, London, UK

Key words

mental ill health screening, military, PTSD, alcohol

Correspondence

Howard Burdett, King's Centre for Military Health Research, Psychological Medicine Department (WEC), c/o 125 Cold Harbour Lane, London SE5 9NU, UK.
Email: howard.burdett@kcl.ac.uk

Received 24 February 2015;
revised 15 September 2015;
accepted 20 October 2015

Abstract

Two-phase mental health screening methods, in which an abridged mental health measure is used to establish who should receive a more comprehensive assessment, may be more efficient and acceptable to respondents than a stand-alone complete questionnaire. Such two-phase methods are in use in US armed forces post-deployment mental health screening. This study assesses the sensitivity and specificity of abridged instruments (used in the first phase) compared to the full instruments (the second phase), and whether false negative cases resulting from the use of abridged tests were detected by another test, among a UK military screening sample. Data from a group of UK Armed Forces personnel ($n = 1464$) who had completed full questionnaires assessing symptoms of post-traumatic stress disorder (PTSD) (PTSD Checklist – Civilian Version, PCL-C) and alcohol misuse (Alcohol Use Disorder Identification Test, AUDIT) were used. An abridged version of the PCL-C performed well in discriminating potential PTSD cases (as measured by the full instrument); AUDIT showed less discriminatory power, particularly due to poor specificity. Many cases missed by one abridged test would have been detected by an alternative test. Thus two-phase screening designs reduce the resource burden of a project without substantial loss of sensitivity for PTSD, but are less effective in discriminating potential cases of alcohol misuse. Copyright © 2016 John Wiley & Sons, Ltd.

Introduction

Two-phase surveys use an initial brief questionnaire to identify participants who may need to be assessed with a more comprehensive measure. Resulting data can be used in epidemiological studies to efficiently estimate

prevalence of mental health and behavioural disorders (Dunn *et al.*, 1999). Two-phase surveys may also be used to screen large populations to decide whether a person would benefit from specialized health services subject to more comprehensive assessment.

The rationale for using a two-phase assessment is multi-fold: shorter questionnaires may improve response rates (Dillman *et al.*, 1993; VanGeest *et al.*, 2007), response quality (Herzog and Bachman, 1981; Galesic and Bosnjak, 2009) and increase efficiency thus reducing costs (Dunn *et al.*, 1999). Better disposition of participants towards the study may reduce the rate of incomplete responses in the questionnaire and reduce questionnaire fatigue as longer questionnaires tend to have lower response rates (Rona *et al.*, 2004). In a two-phase study, the longer questionnaire will be completed only by those who may have probable psychiatric disorders; the interest to complete a questionnaire among those participants may be greater.

A two-phase approach to screening has been applied in a military context: shortly after commencement of the ground war in Iraq in 2003, the US Department of Defense instituted a screening programme whereby military personnel returning from deployment completed a brief post-deployment assessment (the Post-Deployment Health Assessment), which included several short mental health measures. Screening positive for any of these measures (the first phase) identifies an individual as potentially having a mental health or behavioural disorder; personnel who screen positive were interviewed by a health care professional (the second phase), who determines whether further referral for therapy or treatment is required (Deployment Health Clinical Center, 2014). Post-deployment mental health screening is performed by other nations including Australia and Canada (Searle *et al.*, 2015; Zamorski *et al.*, 2014).

The King's Centre for Military Health Research (KCMHR) has been engaged in a randomized controlled trial to determine the effectiveness of mental health screening among UK Armed Forces personnel returning from deployment (the Post-Operational Screening Trial, henceforth "POST"). Baseline data collection in this study involved a two-phase design; participants answered a subset of questions from the full instruments [which included measures for post-traumatic stress disorder (PTSD), depression, anxiety and alcohol misuse]. Respondents were then directed to the full instrument (or not) depending on their scores for that initial subset of questions. In this article the primary aim is to determine how many mental ill health cases identified using the full measures are missed when using the brief measures (i.e. the number of false negatives). By comparing the responses given to the full measure with responses to the subset of questions comprising the shorter measure, it is possible to determine what proportion of those who were not directed to the second phase would have

qualified as a possible case. A similar but independent sample of UK military personnel who had returned from deployment, and completed the same measures in full (the "BATTLEMIND" study), was used as a basis for the estimates.

The overall aim of this article is to examine the consequences of using two-phase instruments, primarily in a mental ill health screening context. This article utilizes data from this UK BATTLEMIND study (Mulligan *et al.*, 2012) to determine: (1) the sensitivity and specificity of a number of short instruments (used in the initial phase) compared to their respective full instruments (the second phase); (2) whether cases missed by one test are detected by another test (and hence nonetheless directed to care in a screening context), additionally, to investigate consequences of using screening-related cutoffs on estimates of prevalence; (3) by extrapolation from the impact of two-phase screening on the BATTLEMIND study we estimate the true prevalences of mental health difficulties in the POST sample.

Methods

The BATTLEMIND sample

The UK "BATTLEMIND" study was a two-arm cluster randomized controlled trial of an interactive mental health education-based intervention conducted among UK Armed Forces who had recently returned from deployment to Afghanistan (Mulligan *et al.*, 2012). The UK BATTLEMIND study collected data at two time points; baseline data was collected on the first day following extraction from the operational area, and during a follow-up period 4–6 months after return to the UK. Randomization occurred at the company level (a unit of around 100 personnel). Of the 2443 individuals sampled at baseline, 66.1% ($n = 1616$) responded at follow-up. This follow-up sample, in which respondents filled full versions of the instruments of interest, is used as the source of data for comparison to the POST sample (described in the next section). This article uses data from the follow-up phase, as the questionnaire at follow-up included the mental health measures used in the POST study. Combined data from both arms of BATTLEMIND is used in this study [there were no significant differences between arms for either the PTSD Checklist (PCL) or the Alcohol Use Disorder Identification Test (AUDIT)]. In order to match the characteristics of the POST sample, only data collected from regular members of the Army and Royal Marines who were still in service at the time of questionnaire completion were used ($n = 1489$, 92% of follow-up sample).

The POST sample

The POST study is a randomized controlled trial, clustered at the platoon level (usually 15–35 individuals) (Figure 1), aiming to determine whether post-deployment screening can detect mental health disorder and promote appropriate help-seeking among military personnel. Data in this article consist of the baseline measures collected 6–12 weeks after the respondents had returned from deployment to Afghanistan. Of a sample of 10,190 personnel, 8726 (85.6%) completed the baseline assessment. The study design used a digital methodology which permitted the use of two-phase questionnaires. By completing questionnaires embedded in a bespoke programme, respondents were directed to each full measure only when their scores indicated possible mental health or alcohol caseness in the first phase. It is intended that the consequences of this two-phase approach be determined by comparison with the BATTLEMIND sample

Measures

This study considered three mental health instruments: the PTSD Checklist – Civilian Version (PCL-C), the AUDIT, and the nine-item Patient Health Questionnaire (PHQ-9). These measures focus on the most common challenges to the mental health of UK Armed Forces personnel (Hotopf *et al.*, 2006; Fear *et al.*, 2010), and are representative of those used in the post-deployment screening processes of other nations: the United States utilizes a short version of the PCL-C (the PC-PTSD), with an additional question adapted from PHQ-9; Australia uses the PCL-C, AUDIT, and Kessler Psychological Distress Scale; and Canadian forces use the

PCL-C and portions of the PHQ (from which the PHQ-9 is derived).

Post-traumatic stress disorder: PCL-C

The PCL-C is a fully validated measure used to detect symptoms of PTSD (Weathers *et al.*, 1994) and which is widely used in military populations in both the United States (Hoge *et al.*, 2004) and UK (Hotopf *et al.*, 2006; Fear *et al.*, 2010). It comprises 17 items with responses indicating degree of severity (coded 1–5, with answers from “not at all” to “extremely”). It has been demonstrated to have a sensitivity of 0.78 and specificity of 0.86 at a PCL cutoff score of 50 when compared to the gold standard Clinician-Administered PTSD Scale (CAPS) (Blake *et al.*, 1990; Blanchard *et al.*, 1996), and a weighted sensitivity of 0.54 and specificity of 0.93 in a systematic assessment of studies utilizing a variety of diagnostic interview methods as gold standards (Terhakopian *et al.*, 2008).

A cutoff of 50 is frequently used to classify respondents as having probable PTSD (Schwartz *et al.*, 1997; Kang *et al.*, 2003; Hotopf *et al.*, 2006; Fear *et al.*, 2010). This cutoff may be appropriate in epidemiological studies where the priority is to gain the most accurate estimate of prevalence – this is found when the number of false positives is as close as possible to the number of false negatives, such that the two numbers effectively cancel. Where the true prevalence is low, this tends to favour specificity over sensitivity (as a small proportion of the larger unaffected population falsely identified as cases can be equal in size to a large proportion of missed cases among the affected minority). Lower cutoffs are more appropriate in some circumstances [e.g. where prevalence is relatively low in the population examined, and/or where higher sensitivity is required despite the loss of specificity (Lang *et al.*, 2003; Cook *et al.*, 2005; Bliese *et al.*, 2008; Terhakopian *et al.*, 2008)]. In a screening context, the priority is to identify potential cases and direct them to care; hence, predictive value, and in particular sensitivity, is more important when using measures for screening.

Studies among the US military have found low sensitivity at the high cutoff of 50; in one, comparison to a structure clinical interview have sensitivity of 0.24 at this cutoff (though high specificity of 0.98), with overall performance at low cutoffs of 30 to 34 (which showed sensitivity from 0.78 to 0.71 with specificity 0.88–0.91) (Bliese *et al.*, 2008). Similarly, a study in the Australian military (validating against a structured diagnostic interview detecting 30-day disorder) found sensitivity and specificity

Figure 1. Design of controlled trial cluster randomized at platoon level.

of 0.30 and 0.97, respectively, at PCL-C cutoff 50, with an optimal screening cutoff (i.e. one which maximized the total of sensitivity and specificity) of 29, for which sensitivity and specificity were 0.79 and 0.78, respectively (Searle *et al.*, 2015). Functional impairment has been demonstrated at levels below the cutoff of 50 (Rona *et al.*, 2009). As the prevalence of PTSD in the UK military is lower than in the United States (Levin, 2014; Sundin *et al.*, 2014), and a higher sensitivity is advisable even at the cost of some loss of specificity, this study used a cutoff score of 40 to indicate the presence of probable PTSD.

Respondents in the UK BATTLEMIND study sample answered the full 17-item PCL-C immediately post-deployment and at follow-up. To simulate BATTLEMIND responses as they would be in the POST study, eight items were used in the initial phase. Items were extracted and scored in pairs; endorsing either of the paired items at the “moderate to extreme” level counted as a score of one for each of the four pairs (thus generating potential first phase scores 0–4). Questions were paired as follows:

- (1) “Repeated, disturbing memories, thoughts, or images of a stressful experience from the past”/“Repeated, disturbing dreams of a stressful experience from the past”
- (2) “Avoid thinking about or talking about a stressful experience from the past or avoid having feelings related to it”/“Avoid activities or situations because they remind you of a stressful experience from the past”
- (3) “Feeling distant or cut off from other people”/“Feeling emotionally numb or being unable to have loving feelings for those close to you”
- (4) “Being ‘super alert’ or watchful on guard”/“Feeling jumpy or easily startled”

These paired questions are analogous to the four-question initial phase of the POST study (following US post-deployment screening protocol), which used the Primary Care PTSD Scale (Prins *et al.*, 2004) comprised of the following yes/no questions from the stem “in your life, have you ever had any experience that was so frightening, horrible, or upsetting that, in the past month, you”:

- (1) Have had nightmares about it or thought about it when you did not want to?
- (2) Tried hard not to think about it or went out of your way to avoid situations that reminded you of it?
- (3) Were constantly on guard, watchful, or easily startled?
- (4) Felt numb or detached from others, activities, or your surroundings?

Respondents in the POST study were directed to the second phase if they responded “yes” to two or more questions in the initial phase.

Alcohol misuse: AUDIT

The AUDIT is a 10-item measure of alcohol misuse (Babor *et al.*, 2001), and has been used in a UK military context (Hotopf *et al.*, 2006; Fear *et al.*, 2010). The 10 items include three questions on hazardous alcohol use (regarding volume and frequency of drinking), three questions representing alcohol dependence, and four questions reflecting potentially harmful alcohol use (including guilt, blackouts, and alcohol-related injuries). Each question is scored 0–4. This study identifies cases of alcohol use potentially harmful to health (alcohol misuse) as having a total score of 16 or more. A cutoff of 20 was also investigated; this was the cutoff of interest in the POST study, as this cutoff is associated with functional impairment in the UK military population, whereas lower scores are not (Rona *et al.*, 2010). A score of eight is the usual cutoff, and has been validated against a number of diagnostic measures, but shows highly variable sensitivity and specificity dependent on the nature of and prevalence in the underlying sample (Reinert and Allen, 2002, 2007); this value is too common to be of value among the UK Armed Forces, which show a high prevalence of alcohol misuse (Fear *et al.*, 2007). Given the difference in alcohol usage and drinking cultures of different forces, international comparisons with other screening programmes with respect to AUDIT are problematic; for reference, however, among the Australian Defence Force the optimal epidemiological cutoff was 20, while the optimal screening cutoff was eight (Searle *et al.*, 2015).

A short instrument was produced from the first two questions (how often the respondent drinks, and how much they drink on a typical day) for the first phase (range 0–8). This was selected for pragmatic reasons – the original protocol for the POST study involved using the first two items from AUDIT and two from the US Post-Deployment Health Assessment (PDHA) alcohol measure; following piloting, the questions from the US PDHA were not found to improve the measure so were removed. In the POST study, respondents were directed to the full AUDIT instrument if they scored six or more on this short instrument.

Depression: PHQ-9

Depression was examined using the validated PHQ-9 (Kroenke *et al.*, 2001; Kroenke *et al.*, 2010). Depression was defined by the endorsement of five or more problems

at a frequency of “more than half the days” (or the participant endorsing having thoughts that he would be better off dead or hurting himself on at least “several days”); these five items must include at least one of “little interest or pleasure in doing things” and/or “feeling down, depressed or hopeless”. Validation studies and meta-analyses indicated a sensitivity of 0.77 to 0.88 and specificity of 0.88 to 0.92 (Kroenke *et al.*, 2010).

The shorter instrument is formed from the first two items of the PHQ-9, which were how often, in the last two weeks, the respondent had experienced “little interest or pleasure in doing things” and “feeling down, depressed or hopeless”. This measure was equivalent to the established short version of PHQ-9, the PHQ-2 (Kroenke *et al.*, 2003), except that the stand alone PHQ-2 is usually used to provide a score from zero to six, whereas here a case is defined as endorsing either item (or both) as “more than half the days” or “nearly every day” (selected as a requirement for caseness in the PHQ-9 is that at least one of these two items is positive). As scoring positive for at least one of these questions is a requirement to qualify as a case under the PHQ-9 it is not possible for this two-phase approach to miss any cases. Consequently, the PHQ-9 should detect cases which would otherwise be missed by other instruments; this study does not seek to test validity of the short instrument to the full instrument, as it has 100% sensitivity by definition.

Analysis

The sensitivity of the short measures of PCL-C and AUDIT against caseness in the full measures was assessed using data from the UK BATTLEMIND sample. The questions which comprised the initial phase of the POST study (as defined earlier) were extracted from UK BATTLEMIND sample data and a cutoff score representing potential caseness was generated. Each possible short measure score was treated as a potential cutoff for comparison with caseness under the full instrument; thus, sensitivity for a range of scores could be calculated. Specificity was also included to determine the value of performing a two-phase method in this context. From these, standard receiver operating characteristic (ROC) curves (Zweig and Campbell, 1993) were plotted, and Youden's Indices, which summarize the performance of a diagnostic test as a single statistic where zero indicates no utility and one indicates infallibility (i.e. sensitivity + specificity – 1) (Youden, 1950) were calculated as a measure of informedness at each initial phase cutoff value.

Those cases in the BATTLEMIND sample (all of whom completed the full measures) which would have been

missed by a two-phase screening approach (as used in the POST sample) for a given measure (PCL-C or AUDIT) were then examined to determine whether these cases would have been detected by the other study questionnaires (including PHQ-9). This was performed assuming short instrument cutoffs were as used in the POST study; additionally, higher short measure cutoff values (i.e. less sensitive but more specific) were tested to see if the resulting lower sensitivity for each separate instrument could be compensated for by using multiple measures to identify individuals with probable mental illness.

From the findings in the sensitivity and specificity analysis in the BATTLEMIND sample, it is possible to extrapolate the “true” prevalence within the POST sample for each instrument, i.e. the prevalence that would have been found by a one-phase rather than two-phase method. The sensitivity measured represents the proportion which would have been detected by the two-phase screening in the POST sample; thus the “true” prevalence can be estimated by multiplying this proportion by the observed prevalence. “True” prevalence is defined in relation to the PCL and AUDIT questionnaires.

Analyses were performed using the statistical software package, STATA (StataCorp, 2009) version 11.

Results

Demographics

Table 1 shows the military demographic characteristics of the two samples in this study. The samples are similar in most regards, except that the BATTLEMIND sample contains a higher proportion of Royal Marines and Commissioned Officers.

Sensitivity and specificity of short versus full instruments

PCL-C

Responses to the full PCL-C among the BATTLEMIND sample had a mean score of 23.7 (standard deviation 9.7), with 8.4% scoring above the cutoff of 40. Table 2 presents the sensitivity and specificity of the short version of PCL-C when detecting PCL-C cases under the full instrument. The cutoff score of two (used in the POST study) offered high sensitivity and specificity and maximizing Youden's Index. Higher specificity can be achieved with higher cutoff values, but at the cost of markedly lower sensitivity. The area under the curve (AUC) (Figure 2) demonstrates the short questionnaire has good discriminating power for detecting cases under the full questionnaire.

Table 1. Military and demographic characteristics of the BATTLEMIND and POST samples¹

Category	Characteristic	BATTLEMIND sample (N= 1489)	POST sample (N= 8725)
		n (%)	n (%)
Sex	Male	1464 (98.3)	8474 (97.1)
	Female	25 (1.7)	251 (2.9)
Rank	Private-equivalent rank	622 (41.8)	3709 (50.4)
	Non-commissioned officers	689 (46.3)	3276 (44.5)
	Commissioned officers	178 (12.0)	375 (5.1)
Service arm	Army	891 (59.8)	7535 (86.4)
	Royal Marines	598 (40.2)	1191 (13.7)
	<25	609 (40.9)	3791 (43.6)
	25–29	424 (28.5)	2535 (29.2)
Age (years)	30–34	201 (13.5)	1365 (15.7)
	35–39	178 (12.0)	706 (8.1)
	40+	77 (5.2)	295 (3.4)

¹Totals within categories may not sum to overall total due to missing data.

Table 2. Sensitivity and specificity of the short PTSD questionnaire (four items) for caseness under the full PCL-C (17 items) with a cutoff score of 40

Short measure score	Number of cases in the four item questionnaire/ number of cases in the 17 item questionnaire (n/N)	Sensitivity (%)	95% Confidence interval (CI) (%)	Non-cases in the four item questionnaire/ non-cases in the 17 item questionnaire (n/N)	Specificity (%)	95% CI (%)	Correctly classified (%)	Youden's Index
0	124/124	100%	97.1–100	0/1348	0	0–0.3	8.4	0
1	124/124	100	97.1–100	1051/1348	78.0	75.7–80.2	79.8	0.78
2	122/124	98.4	94.3–99.8	1256/1348	93.2	91.7–94.5	93.6	0.92
3	99/124	79.8	71.7–86.5	1327/1348	98.4	97.6–99.0	96.9	0.78
4	62/124	50.0	40.9–59.1	1346/1348	99.9	99.5–100	95.7	0.50

AUDIT

Responses to the full AUDIT instrument by BATTLEMIND respondents had a mean score of 11.0 (standard deviation 6.0), with 20.5% scoring above the cutoff of 16 and 9.9% scoring above 20. The sensitivity of the short alcohol misuse questionnaire is very high up to a score of five, decreases somewhat for a score of six and seven, and markedly decreases for a score of eight. Specificity is poor up to a score of five to six, intermediate for a score of seven and high for a score of eight (Table 3 and Figure 3). The AUC of 0.82 indicates that the short questionnaire has weaker discriminating power than the short version of the PCL used. Youden's Index would indicate an optimal cutoff of seven; however, 20–30% of cases would be missed at this cutoff, and a more sensitive cutoff of six may be preferable in a screening context.

Detection of cases by other instruments

At the cutoffs used in the POST study, both of the two cases missed by the short PCL-C (using a first-phase cutoff of two) would have been detected by the short version of PHQ-9 instruments, of which one would be detected by the full instrument. The short version of AUDIT would also have detected both false negatives (except short AUDIT with cutoff seven, which would detect one); neither case would have been detected by the full version of AUDIT.

As regards AUDIT at caseness threshold 16, 10 of the 19 alcohol misuse false negatives (using a first-phase cutoff of six) would have been detected by the other short tests, and nine by the full tests (Figure 4). Increasing the short AUDIT cutoff to seven would result in missing 81 AUDIT cases, of which 39 would be detected by short instruments and 27 by full instruments. Reducing the short AUDIT

Figure 2. Receiver operating characteristic (ROC) curve for first stage against full PCL at cutoff 40.

cutoff to five would result in only six cases being missed, of which four are detected by other study instruments; however, the price for this high sensitivity is that 771 of the 1152 non-cases would have been incorrectly directed to complete the full measure thus greatly decreasing the efficiency of the two phase approach.

For AUDIT at caseness threshold 20, seven cases were missed by the short version of AUDIT (at cutoff six), of which five were detected by other short and full instruments. Increasing the short AUDIT cutoff to seven resulted in 28 missed cases, of which 17 were detected by the other short instruments and 16 by full instruments. Reducing the short AUDIT cutoff score to five reduced

the number of missing cases to three, all of which were detected by other instruments, but again this would carry a high penalty due to poor specificity (921 of the 1305 non-cases would have been directed to the full instrument).

Implications for prevalence

Using the sensitivity values calculated in Tables 2 and 3, estimates of the “true” prevalence values for probable PTSD and alcohol misuse in the POST study were calculated. A one phase approach is estimated to have resulted in a “true” prevalence of 4.9% [95% confidence interval (CI) 4.4–5.3%] for probable PTSD, rather than the 4.8% (95% CI 4.3–5.2%) measured in the POST sample. “True” alcohol misuse at an AUDIT cutoff of 16 is estimated to be 17.6% (95% CI 16.8–18.4%), rather than the 16.5% (95% CI 15.7–17.3%) measured; for the AUDIT cutoff of 20 used in the POST study, the “true” prevalence would be 8.6% (95% CI 8.0–9.2%) rather than the 8.2% (95% CI 7.6–8.7%) measured.

Discussion

The intention of this study was to determine the suitability and efficiency of a two-phase screening methodology, as compared with a single phase. We found that the two-phase approach used in the POST study is unlikely to miss PTSD cases based on analysis of the UK BATTLEMIND study data. However, a short questionnaire to assess

Table 3. Sensitivity and specificity of short measure of AUDIT for detecting alcohol misuse cases using full AUDIT questionnaire at cutoffs 16 and 20. Short measure cutoffs 0–3 not shown; specificity for these cutoffs is too low to be informative.

Short measure score	Number of cases detected in the short questionnaire/number of total case based on full AUDIT (n/N)	Sensitivity (%)	95% Confidence interval (CI) (%)	Non-cases detected short questionnaire/non-cases detected using full AUDIT questionnaire (n/N)	Specificity (%)	95% CI (%)	Correctly classified (%)	Youden's Index
AUDIT at cutoff 16								
4	294/297	99.0	97.1–99.8	220/1152	19.1	16.–921.5	35.5	0.18
5	291/297	98.0	95.7–99.3	381/1152	33.1	30.4–35.9	46.4	0.31
6	278/297	93.6	90.2–96.1	580/1152	50.4	47.4–53.3	59.2	0.44
7	216/297	72.7	67.3–77.7	877/1152	76.1	73.6–78.6	75.4	0.49
8	62/297	20.9	16.4–26.0	1138/1152	98.8	98.0–99.3	82.8	0.20
AUDIT at cutoff 20								
4	142/144	98.6	95.1–99.8	221/1305	16.9	14.9–19.1	25.1	0.16
5	141/144	97.9	94.0–99.6	384/1305	29.4	27.0–32.0	36.2	0.27
6	137/144	95.1	90.2–98.0	592/1305	45.4	42.6–48.1	50.3	0.45
7	116/144	80.6	73.1–86.7	930/1305	71.3	68.7–73.7	72.2	0.52
8	46/144	31.9	24.4–40.2	1275/1305	97.7	96.7–98.4	91.2	0.30

Figure 3. Receiver operating characteristic (ROC) curves for short version of AUDIT against full AUDIT caseness at cutoffs (A) 16 and (B) 20.

alcohol misuse in a two-phase approach only detected 94–95% of cases. Some of these false negative cases for alcohol misuse would have been detected by other measures included in a screening programme, but many would be missed (e.g. 10 of the 19 missed AUDIT-16 cases would have been missed by other instruments, using a short

AUDIT cutoff of six). The prevalence of probable PTSD found by the two-phase process is likely to be accurate, but it is likely that the prevalence of alcohol misuse would have been slightly higher if a one-phase approach were used.

This two-phase approach was used in the POST study to facilitate response to a digital questionnaire but we were concerned that such an approach may cause the loss of sensitivity. The cutoff for the four item PTSD questionnaire used in the POST study appeared to have the greatest utility; any other cutoff would have reduced performance. The cutoff of six for the shorter AUDIT questionnaire used in the first phase of the POST study was more equivocal. A score of five in the short questionnaire would not have offered clear advantages, as the sensitivity would have been acceptable but the price too high in terms of a low specificity. In contrast a score of seven would have provided a more reasonable specificity, and maximized Youden's Index, but has poor sensitivity and many cases are not detected by other instruments. Thus, in the context of a screening study where sensitivity is usually more important than specificity, a cutoff of six is preferred.

The weaker discriminatory power of the AUDIT study may be due to the fact that the two questions asked (which are common to most versions of abridged AUDIT questionnaires) are on frequency and volume of drinking, while later questions include alcohol harm and dependence. In a population where heavy drinking is a cultural norm such as the UK military, these questions may be insufficient to discriminate heavy drinkers who perceive themselves to be high functioning. We would argue that, despite the poorer performance of the shorter version of the AUDIT than other tests of mental illness in the POST study, the advantages would be greater than using the full AUDIT questionnaire. The reason for this is that if we had used a short questionnaire for all other tests to detect

Figure 4. Detection of AUDIT-16 cases by other instruments that were missed by the short AUDIT instrument.

possible mental illness (except alcohol misuse), participants would have believed that the main reason for the study was to ascertain cases of alcohol misuse as proportionally more questions on alcohol would appear in a questionnaire purporting to be a short questionnaire. This could increase unwillingness to cooperate with the screening programme given that high alcohol consumption is widespread in the UK military and service members are less willing to recognize a problem with alcohol than for other mental ill health conditions (Hines *et al.*, 2014). An alternative short AUDIT measure may be more appropriate for two-phase screening.

As most cases which were missed by the initial test were detected by other instruments, sensitivity may in practice be higher than it appears. However, health carers need to be aware that within a screening programme it is possible that other conditions may be a concern in addition to that identified by the screening questionnaire. To make full use of this finding in future studies, we suggest that screening programmes use multiple measures when attempting to detect mental health disorder and alcohol misuse.

Strengths and weaknesses

This study utilized one sample of UK personnel who had returned from deployment (the UK BATTLEMIND sample), and extrapolated findings to the sample of interest (the POST sample). Using an independent sample allows these findings to be more generally applicable, and avoids the necessity of splitting a dataset and hence reducing analytical power. Although the BATTLEMIND sample assessed personnel at a slightly different post-deployment time than the POST sample, and there may be differences in prevalence between the two samples, there is no reason to assume significant differences in the sensitivity and specificity of the short tests in relation to the full instruments between the two samples.

Two-phase screening methods for mental health disorders typically use standardized clinical interviews as the gold standard for comparison; this sample instead

uses the full instrument as the equivalent of a “gold standard”. This is appropriate given that the context of this study is restricted to the use of self-report questionnaires. Also, questionnaire-based estimates of prevalence are affected by the cutoff used and the underlying prevalence, and may not correlate well with clinical diagnoses obtained at interview (Terhakopian *et al.*, 2008). However, it has been shown that all the tests used in our screening programme for probable PTSD and alcohol misuse have high validity.

Conclusions

Our findings demonstrate that a two phase screening questionnaire using short instruments is appropriate for increasing efficiency without a potential loss of cases of possible mental illness. Although there may be a small proportion of missed alcohol misuse cases, most individuals with probable PTSD would have been detected by such an approach.

Acknowledgements

This study was funded by the US Congressionally Directed Medical Research Programmes Award Number W81XWH-10-1-0881. Some authors of this paper are funded by other grants. Financial support was also received from the Department of Health via the National Institute for Health Research (NIHR) Comprehensive Biomedical Research Centre Award to Guy's & St Thomas' National Health Service Foundation Trust in partnership with King's College London and King's College Hospital National Health Service Foundation Trust.

Declaration of interest statement

At the time of the BATTLEMIND study, and at the time that the POST study was conceived, NG was a member of the UK Armed Forces but has subsequently retired. NJ is a serving member of the UK Armed Forces. SW is honorary civilian consultant advisor in psychiatry to the British Army.

References

- Babor T.F., Higgins-Biddle J.C., Saunders J.B., Monteiro M.G. (2001) *AUDIT. The Alcohol Use Disorders Identification Test*, Geneva: Department of Mental Health and Substance Dependence, World Health Organization.
- Blake D., Weathers F., Nagy L., Kaloupek D., Klauminzer G., Charney D., et al. (1990) A clinician rating scale for assessing current and lifetime PTSD: the CAPS-1. *The Behavior Therapist*, **13**, 187–188.
- Blanchard E.B., Jones-Alexander J., Buckley T.C., Forneris C.A. (1996) Psychometric properties of the PTSD Checklist (PCL). *Behaviour Research & Therapy*, **34**(8), 669–673.
- Bliese P.D., Wright K.M., Adler A.B., Cabrera O., Castro C.A., Hoge C.W. (2008) Validating the primary care posttraumatic stress disorder screen and the posttraumatic stress disorder checklist with soldiers returning from combat. *Journal of Consulting and Clinical Psychology*, **76**(2), 272.
- Cook J.M., Elhai J.D., Areán P.A. (2005) Psychometric properties of the PTSD Checklist with older primary care patients. *Journal of Traumatic Stress*, **18**(4), 371–376.
- Deployment Health Clinical Center. (2014) Enhanced Post-Deployment Health Assessment (PDHA) Process (DD Form 2796).

- https://www.pdhealth.mil/dcs/dd_form_2796.asp [7 October 2014].
- Dillman D.A., Sinclair M.D., Clark J.R. (1993) Effects of questionnaire length, respondent-friendly design, and a difficult question on response rates for occupant-addressed census mail surveys. *Public Opinion Quarterly*, **57**(3), 289–304.
- Dunn G., Pickles A., Tansella M. (1999) Two-phase epidemiological surveys in psychiatric research. *The British Journal of Psychiatry*, **174**(2), 95–100.
- Fear N.T., Iversen A., Meltzer H., Workman L., Hull L., Greenberg N. et al. (2007) Patterns of drinking in the UK Armed Forces. *Addiction*, **102**(11), 1749–1759.
- Fear N.T., Jones M., Murphy D., Hull L., Iversen A.C., Coker B. et al. (2010) What are the consequences of deployment to Iraq and Afghanistan on the mental health of the UK armed forces? A cohort study. *The Lancet*, **375**(9728), 1783–1797.
- Galesic M., Bosnjak M. (2009) Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, **73**(2), 349–360.
- Herzog A.R., Bachman J.G. (1981) Effects of questionnaire length on response quality. *Public Opinion Quarterly*, **45**(4), 549–559.
- Hines L.A., Goodwin L., Jones M., Hull L., Wessely S., Fear N.T. et al. (2014) Factors affecting help seeking for mental health problems after deployment to Iraq and Afghanistan. *Psychiatric Services*, **65**(1), 98–105.
- Hoge C.W., Castro C.A., Messer S.C., McGurk D., Cotting D.L., Koffman R. (2004) Combat duty in Iraq and Afghanistan, mental health problems, and barriers to care. *New England Journal of Medicine*, **351**(1), 13–22.
- Hotopf M., Hull L., Fear N.T., Browne T., Horn O., Iversen A. et al. (2006) The health of UK military personnel who deployed to the 2003 Iraq war: a cohort study. *Lancet*, **367**(9524), 1731–1741.
- Kang H.K., Natelson B.H., Mahan C.M., Lee K.Y., Murphy F.M. (2003) Post-traumatic stress disorder and chronic fatigue syndrome-like illness among Gulf War veterans: a population-based survey of 30,000 veterans. *American Journal of Epidemiology*, **157**(2), 141–148.
- Kroenke K., Spitzer R.L., Williams J.B. (2001) The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, **16**(9), 606.
- Kroenke K., Spitzer R.L., Williams J.B. (2003) The Patient Health Questionnaire-2: validity of a two-item depression screener. *Medical Care*, **41**(11), 1284–1292.
- Kroenke K., Spitzer R.L., Williams J.B., Löwe B. (2010) The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. *General Hospital Psychiatry*, **32**(4), 345–359.
- Lang A.J., Laffaye C., Satz L.E., Dresselhaus T.R., Stein M.B. (2003) Sensitivity and specificity of the PTSD checklist in detecting PTSD in female veterans in primary care. *Journal of Traumatic Stress*, **16**(3), 257–264.
- Levin A. (2014) PTSD, Suicide Rates Differ Among US, UK Troops. http://psychiatryonline.org/doi/10.1176/pn.47.15.psychnews_47_15_7-a
- Mulligan K., Fear N.T., Jones N., Alvarez H., Hull L., Naumann U. et al. (2012) Postdeployment Battlemind training for the UK armed forces: a cluster randomized controlled trial. *Journal of Consulting and Clinical Psychology*, **80**(3), 331.
- Prins A., Ouimette P., Kimerling R., Camerond R. P., Hugheshofer D. S., Shaw-Hegwar J., et al. (2004) The Primary Care PTSD Screen (PC-PTSD): development and operating characteristics. *International Journal of Psychiatry in Clinical Practice*, **9**(1), 9–14.
- Reinert D.F., Allen J.P. (2002) The Alcohol Use Disorders Identification Test (Audit): a review of recent research. *Alcoholism: Clinical and Experimental Research*, **26**(2), 272–279.
- Reinert D.F., Allen J.P. (2007) The Alcohol Use Disorders Identification Test: an update of research findings. *Alcoholism: Clinical and Experimental Research*, **31**(2), 185–199.
- Rona R.J., Jones M., Fear N.T., Hull L., Hotopf M., Wessely S. (2010) Alcohol misuse and functional impairment in the UK Armed Forces: a population-based study. *Drug and Alcohol Dependence*, **108**(1–2), 37–42.
- Rona R.J., Jones M., French C., Hooper R., Wessely S. (2004) Screening for physical and psychological illness in the British Armed Forces: I: The acceptability of the programme. *Journal of Medical Screening*, **11**(3), 148–152.
- Rona R.J., Jones M., Iversen A., Hull L., Greenberg N., Fear N.T. et al. (2009) The impact of posttraumatic stress disorder on impairment in the UK military at the time of the Iraq war. *Journal of Psychiatric Research*, **43**(6), 649–655.
- Schwartz D.A., Doebbeling B.N., Merchant J.A., Barret D.H. (1997) Self-reported illness and health status among Gulf War veterans: a population-based study. *JAMA: Journal of the American Medical Association*, **277**(3), 238–245.
- Searle A.K., Van Hooff M., McFarlane A.C., Davies C.E., Fairweather-Schmidt A.K., Hodson S.E. et al. (2015) The validity of military screening for mental health problems: diagnostic accuracy of the PCL, K10 and AUDIT scales in an entire military population. *International Journal of Methods in Psychiatric Research*, **24**(1), 32–45.
- StataCorp (2009) *Stata Statistical Software*, TX, StataCorp LP: College Station.
- Sundin J., Herrell R.K., Hoge C.W., Fear N.T., Adler A.B., Greenberg N. et al. (2014) Mental health outcomes in US and UK military personnel returning from Iraq. *The British Journal of Psychiatry*, **204**(3), 200–207.
- Terhakopian A., Sinaii N., Engel C.C., Schnurr P.P., Hoge C.W. (2008) Estimating population prevalence of posttraumatic stress disorder: an example using the PTSD checklist. *Journal of Traumatic Stress*, **21**(3), 290–300.
- VanGeest J.B., Johnson T.P., Welch V.L. (2007) Methodologies for improving response rates in surveys of physicians: a systematic review. *Evaluation & the Health Professions*, **30**(4), 303–321.
- Weathers F.W., Litz B., Huska J.A., Keane T. (1994) *PTSD Checklist – Civilian Version*, Boston, MA: National Center for PTSD, Behavioral Science Division.
- Youden W.J. (1950) Index for rating diagnostic tests. *Cancer*, **3**(1), 32–35.
- Zamorski M.A., Rusu C., Garber B.G. (2014) Prevalence and correlates of mental health problems in Canadian Forces personnel who deployed in support of the mission in Afghanistan: findings from postdeployment screenings, 2009–2012. *Canadian journal of psychiatry. Revue canadienne de psychiatrie*, **59**(6), 319–326.
- Zweig M.H., Campbell G. (1993) Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, **39**(4), 561–577.