Alcohol and Alcoholism Vol. 50, No. 2, pp. 244–250, 2015
Advance Access Publication 22 December 2014
doi: 10.1093/alcalc/agu095

Accuracy of Alcohol Use Disorders Identification Test for Detecting Problem Drinking in 18–35 Year-Olds in England: Method Comparison Study

David R. Foxcroft*, Lesley A. Smith, Hayley Thomas and Sarah Howcutt

Department of Psychology, Social Work and Public Health, Oxford Brookes University, Oxford OX3 0FL, UK
*Corresponding author: Department of Psychology, Social Work and Public Health, Oxford Brookes University, Oxford OX3 0FL, UK. Tel.: +44-1865-485283; E-mail: david.foxcroft@brookes.ac.uk

(*Received* 29 *July* 2014; *first review notified* 9 *September* 2014; *in revised form* 14 *November* 2014; *accepted* 3 *December* 2014)

Abstract — Aims: To assess the accuracy of Alcohol Use Disorders Identification Test (AUDIT) scores for problem drinking in males and females aged 18–35 in England. **Methods:** A method comparison study with 420 primary care patients aged 18–35. Test measures were AUDIT and AUDIT-C. Reference standard measures were (a) Time-Line Follow-Back interview for hazardous drinking; World Mental Health Composite International Diagnostic Interview for (b) DSM-IV alcohol abuse, (c) DSM-IV alcohol dependence, (d) DSM-5 alcohol use disorders. **Results:** Area under the curve (AUC) was (a) 0.79 (95% CI 0.73–0.85; males) and 0.84 (0.79–0.88; females); (b) 0.62 (0.54–0.72; males) and 0.65 (0.57–0.72; females); (c) 0.77 (0.65–0.87; males) and 0.76 (0.67–0.74; females); (d) 0.70 (0.60–0.78; males) and 0.73 (CI 0.67–0.78; females). Identification of threshold cut-point scores from the AUC was not straightforward. Youden *J* statistic optimal cut-point scores varied by 4–6 AUDIT scale points for each outcome according to whether sensitivity or specificity were prioritized. Using Bayes' Theorem, the post-test probability of drinking problems changed as AUDIT score increased, according to the slope of the probability curve. **Conclusion:** The full AUDIT scale showed good or very good accuracy for all outcome measures for males and females, except for alcohol abuse which had sufficient accuracy. In a screening scenario where sensitivity might be prioritized, the optimal cut-point is lower than established AUDIT cut-points of 8+ for men and 6+ for women. Bayes' Theorem to calculate individual probabilities for problem drinking offers an alternative to arbitrary cut-point threshold scores in screening and brief intervention programmes.

INTRODUCTION

The European Union (EU) is the heaviest drinking region of the world, drinking 11 l of pure alcohol per adult each year (Anderson and Baumberg, 2006). More than 1 in 4 deaths among men (aged 15–29 years) and 1 in every 10 deaths among young women in the EU is alcohol-related (Rehm *et al.*, 2005). Young people (aged 15–24 years) are responsible for a high proportion of this burden, with over 25% of youth male mortality and ~10% of young female mortality being due to alcohol (Anderson and Baumberg, 2006). Screening for drinking problems in early adulthood has the potential to identify those at risk and trigger an appropriate intervention, whether it is brief advice or onward referral. However, the uncertainty over the accuracy of screening tests for alcohol problems was reflected in the UK NHS National Screening Committee decision in 2011 (Lines, 2010) to recommend against the adoption of a formal universal screening programme for alcohol misuse. They concluded that whilst the Alcohol Use Disorders Identification Test (AUDIT) (Babor *et al.*, 2001) had been evaluated in many studies and there was some UK evidence for AUDIT as a valid test for alcohol misuse in men aged 35–54, it is not clear which threshold scores should be used for women, younger adults, adults over 65 years and ethnic minorities.

The accuracy of diagnostic or screening tests for identifying problem drinkers in primary care settings has been evaluated by systematic review and meta-analysis (Kriston *et al.*, 2008; Smith *et al.*, 2014). The Cochrane review (Smith *et al.*, 2014) found that the AUDIT had good accuracy for identifying alcohol abuse or dependence, and that the short-form AUDIT-C had good accuracy for identification of hazardous drinking. Only one British study ($N = 194$) with men (mean age 46 years) from a Welsh primary care setting was included; there were no studies that predominantly included young adults aged 18–35 years or women from a UK primary care population.

There is a lack of direct evidence for optimal cut-point threshold scores for hazardous drinking, alcohol abuse, alcohol dependence and alcohol use disorders in a younger adult population. Another consideration is whether selecting a specific cut-point threshold, thus dichotomizing the test scale, could be practically inferior to the more detailed information provided with the application of Bayes' Theorem (Foxcroft *et al.*, 2009), where the probability of having an alcohol problem can be estimated for each discrete test score. This approach (Hall, 1985) combines the prior probability (e.g. population prevalence) with the likelihood ratio derived from each test score to create a post-test probability. A probability graph (Katz, 1974) or a nomogram (Fagan, 1975) can then be produced.

In this accuracy study we (a) assessed the accuracy of the AUDIT and AUDIT-C for the detection of hazardous drinking, alcohol abuse (DSM-IV), alcohol dependence (DSM-IV) and alcohol use disorders (DSM-5) with young adults in a UK primary care population using ROC analysis and suggested optimal cut-point threshold scores; and (b) calculated probability estimates for each discrete test score using Bayes' Theorem.

METHOD

This was a method comparison study, with 14 primary care practices in the Thames Valley area of England. For each primary care practice, administrative staff identified patients to be approached for recruitment into the study using a systematic sampling protocol. All eligible patients (aged 18–35) were listed separately from the complete practice population, and the total number of eligible patients (separated by gender) was divided by the target number required in the sample, with adjustments made for expected non-response and refusal. The target number was the same for each practice. This gave the size of step to be applied to the list of eligible patients, and

practice administrative staff selected patients from the list at every step. For example, if the practice had 1500 eligible 18–35 year-old males on the list, and the target number to be approached was 500 males, then every third male patient would be selected. Lists were typically sorted alphabetically. Because of data protection requirements, all sampling and initial contact with patients was undertaken by practice staff and research personnel were not involved. Across the 14 primary care practices included in the study, the number of eligible patients was 41,412, and the number approached for inclusion in the study 14,480 (35%). These patients (6180 men and 8300 women) were sent a 30-item General Lifestyle Questionnaire (GLQ) that included the 10-item AUDIT, to complete and return. All patients, regardless of AUDIT score, were also invited to participate in a telephone interview following return of the questionnaire. Questionnaire return indicated consent to participate in the study. Telephone interviews were conducted by researchers who were blind to AUDIT responses and score, using (a) Time-Line Follow-Back (TLFB) to ascertain quantity and frequency of alcohol consumption in the previous 90 days (Sobell *et al.*, 1988, 1992); and (b) World Mental Health Composite International Diagnostic Interview (WMH-CIDI) (Robins *et al.*, 1988) to assess alcohol abuse, alcohol dependence and alcohol use disorders. Three telephone interviewers were trained and certified in the use of the WMH-CIDI via a World Health Organisation (WHO) authorized WMH-CIDI Training and Reference Centre. There is no formal external training and certification requirement for TLFB, though we consulted with the TLFB developers and interviewers undertook relevant in-house training. The research procedure did not allow for randomization of the order of administration; all participants completed the AUDIT first, then within a target 2-week follow-up period completed the WMH-CIDI followed by the TLFB.

Quantity of alcohol ascertained via TLFB, was standardized into UK units (one unit is 10 ml (7.9 g) pure alcohol). Hazardous drinking was defined as exceeding 14 (women) or 21 (men) units of alcohol in any 1 week; or 2 (women) or 3 (men) units a day for 5 days in any 1 week (Royal College of Physicians, 2011). We defined '1 week' as any rolling consecutive 7-day period. A positive reference test for hazardous drinking was any incidence in the previous 90 days.

Alcohol abuse and alcohol dependence variables (DSM-IV criteria (American Psychiatric Association, 1994)) were created from WMH-CIDI data using algorithms provided by the WHO WMH-CIDI Centre at Harvard University. DSM-IV Criteria for Alcohol Abuse are a maladaptive pattern of alcohol abuse leading to clinically significant impairment or distress, as manifested by one or more of the following, occurring within a 12-month period:

(1) Recurrent alcohol use resulting in failure to fulfil major role obligations at work, school or home (e.g. repeated absences or poor work performance related to substance use; substance-related absences, suspensions or expulsions from school; or neglect of children or household).
(2) Recurrent alcohol use in situations in which it is physically hazardous (e.g. driving an automobile or operating a machine).
(3) Recurrent alcohol-related legal problems (e.g. arrests for alcohol-related disorderly conduct).
(4) Continued alcohol use despite persistent or recurrent social or interpersonal problems caused or exacerbated by the effects of the alcohol (e.g. arguments with spouse about consequences of intoxication or physical fights).

These symptoms must never have met the criteria for alcohol dependence. DSM-IV Criteria for Alcohol Dependence are a maladaptive pattern of alcohol use, leading to clinically significant impairment or distress, as manifested by three or more of the following seven criteria, occurring at any time in the same 12-month period:

(1) Tolerance, as defined by either of the following: a need for markedly increased amounts of alcohol to achieve intoxication or desired effect; or markedly diminished effect with continued use of the same amount of alcohol.
(2) Withdrawal, as defined by either of the following: the characteristic withdrawal syndrome for alcohol (refer to DSM-IV for further details); or alcohol is taken to relieve or avoid withdrawal symptoms.
(3) Alcohol is often taken in larger amounts or over a longer period than was intended.
(4) There is a persistent desire or there are unsuccessful efforts to cut down or control alcohol use.
(5) A great deal of time is spent in activities necessary to obtain alcohol, use alcohol or recover from its effects.
(6) Important social, occupational, or recreational activities are given up or reduced because of alcohol use.
(7) Alcohol use is continued despite knowledge of having a persistent or recurrent physical or psychological problem that is likely to have been caused or exacerbated by the alcohol (e.g. continued drinking despite recognition that an ulcer was made worse by alcohol consumption).

The new DSM-5 (American Psychiatric Association, 2013) alcohol use disorders variable was created from WMH-CIDI data by the authors (code available on request). Alcohol use disorders in DSM-5 combines the DSM-IV categories of abuse and dependence into a single disorder measured on a continuum from mild to severe. Whereas a diagnosis of alcohol abuse previously required only one symptom, mild alcohol use disorder in DSM-5 requires two to three symptoms from a list of 11. In addition, craving has been added to the list, and problems with law enforcement have been dropped because of cultural considerations that make the criteria difficult to apply internationally.

Our sample size calculation indicated that we required 259 women and 139 men for validation of hazardous drinking, if sensitivity was 80% [1], ±10% (Coulton *et al.*, 2006), given expected prevalence (44% for men; 24% for women) (McManus *et al.*, 2009) with 5% alpha.

ROC curves, sensitivity, specificity, positive and negative predictive values and positive likelihood ratios were calculated, along with 95% confidence intervals. Area under the Curve (AUC) values were calculated to give an indication of the usefulness of the tests for various reference standard measures. A 'rule-of-thumb' approach was used to interpret AUC, with values of between 0.5 and 0.6 indicating low accuracy, between 0.6 and 0.7 indicating sufficient accuracy, between 0.7 and 0.8 indicating good accuracy, between 0.8 and

0.9 indicating very good accuracy, and excellent accuracy if over 0.9 (Fischer *et al.*, 2003; Streiner and Cairney, 2007; Simundic, 2008). Unweighted and weighted Youden index scores were also calculated to indicate potential optimal threshold (cut-point) test scores. The Youden index (*J*) (Youden, 1950), a function of sensitivity and specificity, is a commonly used measure of overall diagnostic effectiveness (Wieand *et al.*, 1989; Goddard and Hinbery, 1990; Zweig and Campbell, 1993), and the weighted index (*Jw*) enables greater emphasis to be placed on either sensitivity or specificity (Li *et al.*, 2013). We used Bayes' theorem to calculate post-test probability by reference test, index test and gender. Pre-test probabilities were based on prevalence figures from the obtained sample (by gender), and post-test probability across the range of test scores was estimated by function fitting and bootstrapping as reported in previous work (Foxcroft *et al.*, 2009). All analyses were undertaken using [R] statistical software (R Development Core Team, 2008; Robin *et al.*, 2011). The National Research Ethics Service approved the study (NRES No. 12/SC/0535); all participants provided informed consent prior to their participation in the study.

RESULTS

Data collection took place between January and October 2013. Of the 14,480 patients invited to participate in the study, 1022 (7.1%) patients agreed by returning the GLQ. Of these, 626 (61.3%) also consented to be interviewed. We completed 420 (138 men and 282 women) telephone interviews within our target timeframe of 2 weeks following return of the GLQ. Data collection was stopped at $N = 420$ as we had achieved our target sample size and we ran out of time for conducting more interviews. Therefore 206 consenters were not interviewed; these were either difficult to reach and arrange a suitable time for interview, or they were not needed as data collection was stopped. Comparing our achieved sample ($N = 420$) with Lower Layer Super Output Area Index of Multiple Deprivation (IMD) quintiles for England (2007), most respondents (53%) came from the lowest deprivation quintile; only 10% were from the two highest deprivation quintiles. The majority were white (86%), and 25% were aged 18–24, 32% aged 25–29 and 43% aged 30–35. See Table 1.

Table 1. Sample demographic information, by gender

	Males *n* (%)	Females *n* (%)	Total *n* (%)
Age Group	$n = 138$	$n = 282$	$n = 420$
18–24	35 (25%)	71 (25%)	106 (25%)
25–29	37 (27%)	95 (34%)	132 (32%)
30–35	66 (48%)	116 (41%)	182 (43%)
Ethnicity	$n = 138$	$n = 282$	$n = 420$
White	122 (88%)	240 (85%)	362 (86%)
Other *(categories merged)*	16 (12%)	42 (15%)	58 (14%)
IMD Quintile (England, 2007)	$n = 128$	$n = 269$	$n = 397$
I (lowest deprivation)	67 (52%)	144 (54%)	211 (53%)
II	21 (16%)	52 (19%)	73 (19%)
III	27 (21%)	45 (17%)	72 (18%)
IV	11 (9%)	25 (9%)	36 (9%)
V (highest deprivation)	2 (2%)	3 (1%)	5 (1%)

Using TLFB reference standard data, 49% (67) men and 51% (144) women were classified as hazardous drinkers. Using WMH-CIDI reference standard data, 36% (49) men and 19% (53) women were classified positive for DSM-IV alcohol abuse, 13% (18) men and 8.5% (24) women were classified positive for DSM-IV alcohol dependence, and 52% (72) men and 40% (112) women were classified positive for DSM-5 alcohol use disorders (none vs. mild/moderate/severe).

Table 2 shows and compares the area under the curve (AUC) for the AUDIT and AUDIT-C tests for hazardous drinking in men and women. The AUCs with respective 95% CIs indicate that both tests have good or very good accuracy for the respective reference standard with no evidence of a difference between the two tests for hazardous drinking (males: bootstrap test (*D*) for correlated ROC curves −1.51, $P = 0.13$; females: $D = -1.33$. $P = 0.19$). The AUC for alcohol abuse, alcohol dependence and alcohol use disorders, with 95% CIs, is also shown in Table 1, and indicates that AUDIT is a good or very good accuracy test for dependence and disorders, but less so for abuse (sufficient accuracy).

Table 3 shows the sensitivity, specificity, positive and negative predictive values, and the positive likelihood ratio for optimal threshold scores for AUDIT and AUDIT-C, respectively, according to: (a) Youden *J* index, sensitivity:specificity weighted equally; (b) Youden *J* index weighted 75:25 for sensitivity:specificity; and (c) Youden *J* index weighted 25:75 for sensitivity:specificity. Optimal cut-points for identification of hazardous drinking using AUDIT were nine and four for men and women, respectively. The optimum cut-point decreased to five and two when weighting favoured sensitivity, and increased to eleven and seven when weighting favoured specificity, for men and women, respectively. Optimal cut-points for identification of hazardous drinking using AUDIT-C were five and four for men and women, respectively. The optimum cut-point decreased to four and two when weighting favoured

Table 2. Area under the curve (AUC) for AUDIT and AUDIT-C as predictors of hazardous drinking classification measured using Time-Line Follow-Back (TLFB), and AUDIT for classification of DSM alcohol problems measured using the World Mental Health Composite International Diagnostic Interview (WMH-CIDI), in males and females

	Males ($n = 138$)		Females ($n = 282$)	
Reference standard measure	AUDIT AUC (95% CI)	AUDIT-C AUC (95% CI)	AUDIT AUC (95% CI)	AUDIT-C AUC (95% CI)
TLFB hazardous drinker	0.79 (0.73–0.85)	0.82 (0.76–0.88)	0.84 (0.79–0.88)	0.85 (0.82–0.90)
WMH-CIDI DSM-IV abuse	0.62 (0.54–0.72)	NA	0.65 (0.57–0.72)	NA
WMH-CIDI DSM-IV dependence	0.77 (0.65–0.87)	NA	0.76 (0.67–0.74)	NA
WMH-CIDI DSM-5 disorder	0.70 (0.60–0.78)	NA	0.73 (0.67–0.78)	NA

sensitivity, and increased to seven and six when weighting favoured specificity, for men and women, respectively.

Optimal cut-points for AUDIT as a predictor of DSM alcohol problems are shown in Table 4. Optimal cut-points for identification of DSM-IV alcohol abuse were ten and five for men and women, respectively. The optimum cut-point decreased to five and two when weighting favoured sensitivity, and increased to fifteen and ten when weighting favoured specificity, for men and women, respectively. Optimal cut-points for identification of DSM-IV alcohol dependence were twelve and seven for men and women, respectively. The optimum cut-point decreased to nine and two when weighting favoured sensitivity, and increased to twelve and eleven when weighting favoured specificity, for men and women, respectively. Optimal cut-points for identification of DSM-5 alcohol use disorders (none vs. mild/moderate/severe) were ten and six for men and women, respectively. The optimum cut-point decreased to 5 and 2 when weighting favoured sensitivity, and increased to 13 and 11 when weighting favoured specificity, for men and women, respectively.

Post-test probability curves, for hazardous drinking, DSM-IV alcohol abuse, DSM-IV alcohol dependence, and DSM-5 alcohol use disorders, along the range of AUDIT scores, are shown in Fig. 1.

Table 3. Optimal cut-points and test characteristics for AUDIT and AUDIT-C as a predictor of hazardous drinking in males and females, according to unweighted and weighted Youden *J* statistic

Reference standard—Reference test	Optimal cut-point	Sens (95% CI)	Spec (95% CI)	PPV (95% CI)	NPV (95% CI)	+LR (95% CI)
Males						
Hazardous drinker—AUDIT						
(i) Unweighted Youden *J*	9	0.64 (0.52–0.76)	0.82 (0.71–0.90)	0.77 (0.64–0.87)	0.71 (0.60–0.80)	3.51 (2.08–5.91)
(ii) Youden *J* weighted for sensitivity	5	0.93 (0.83–0.98)	0.48 (0.36–0.60)	0.63 (0.52–0.72)	0.87 (0.73–0.96)	1.78 (1.41–2.24)
(iii) Youden *J* weighted for specificity	11	0.49 (0.37–0.62)	0.92 (0.83–0.97)	0.85 (0.69–0.94)	0.66 (0.55–0.75)	5.83 (2.61–13.01)
Hazardous Drinker ~ AUDIT-C						
(i) Unweighted Youden *J*	5	0.82 (0.71–0.90)	0.69 (0.57–0.79)	0.71 (0.60–0.81)	0.80 (0.68–0.89)	2.65 (1.84–3.82)
(ii) Youden *J* weighted for sensitivity	4	0.94 (0.85–0.98)	0.51 (0.39–0.63)	0.64 (0.54–0.74)	0.90 (0.76–0.97)	1.91 (1.50–2.43)
(iii) Youden *J* weighted for specificity	7	0.52 (0.40–0.65)	0.93 (0.84–0.98)	0.88 (0.73–0.96)	0.67 (0.57–0.76)	7.42 (3.09–17.80)
Females						
Hazardous drinker—AUDIT						
(i) Unweighted Youden *J*	4	0.88 (0.82–0.93)	0.67 (0.59–0.75)	0.74 (0.67–0.80)	0.85 (0.76–0.91)	2.70 (2.11–3.46)
(ii) Youden *J* weighted for sensitivity	2	1.00 (0.96–1.00)	0.36 (0.28–0.45)	0.62 (0.55–0.68)	1.00 (0.90–1.00)	1.57 (1.38–1.78)
(iii) Youden *J* weighted for specificity	7	0.57 (0.48–0.65)	0.91 (0.84–0.95)	0.86 (0.78–0.93)	0.67 (0.60–0.74)	6.04 (3.53–10.34)
Hazardous drinker—AUDIT-C						
(i) Unweighted Youden *J*	4	0.82 (0.75–0.88)	0.75 (0.67–0.82)	0.78 (0.70–0.84)	0.80 (0.72–0.86)	3.33 (2.46–4.50)
(ii) Youden *J* weighted for sensitivity	2	1.00 (0.96–1.00)	0.37 (0.29–0.46)	0.62 (0.56–0.69)	1.00 (0.90–1.00)	1.59 (1.40–1.80)
(iii) Youden *J* weighted for specificity	6	0.47 (0.39–0.56)	0.95 (0.90–0.98)	0.91 (0.82–0.96)	0.63 (0.56–0.70)	9.31 (4.43–19.55)

Table 4. Optimal cut-points and test characteristics for AUDIT as a predictor of alcohol abuse, alcohol dependence and alcohol use disorders in males and females, according to unweighted and weighted Youden *J* statistic

Reference standard	Optimal cut-point	Sens (95% CI)	Spec (95% CI)	PPV (95% CI)	NPV (95% CI)	+LR (95% CI)
Males						
DSM-IV abuse						
(i) Unweighted Youden *J*	10	0.49 (0.34–0.64)	0.74 (0.64–0.83)	0.51 (0.36–0.66)	0.73 (0.62–0.81)	1.90 (1.20–2.98)
(ii) Youden *J* weighted for sensitivity	5	0.80 (0.66–0.90)	0.33 (0.23–0.43)	0.39 (0.30–0.50)	0.74 (0.58–0.87)	1.18 (0.96–1.45)
(iii) Youden *J* weighted for specificity	15	0.18 (0.09–0.32)	0.90 (0.82–0.95)	0.50 (0.26–0.74)	0.67 (0.57–0.75)	1.82 (0.77–4.27)
DSM-IV dependence						
(i) Unweighted Youden *J*	12	0.67 (0.41–0.87)	0.86 (0.78–0.92)	0.41 (0.24–0.61)	0.94 (0.88–0.98)	4.71 (2.72–8.14)
(ii) Youden *J* weighted for sensitivity	9	0.78 (0.52–0.94)	0.65 (0.56–0.73)	0.25 (0.14–0.38)	0.95 (0.88–0.99)	2.22 (1.57–3.14)
(iii) Youden *J* weighted for specificity	12	0.67 (0.41–0.87)	0.86 (0.78–0.92)	0.41 (0.24–0.61)	0.94 (0.88–0.98)	4.71 (2.72–8.14)
DSM-5 disorders						
(i) Unweighted Youden *J*	10	0.48 (0.35–0.60)	0.78 (0.67–0.87)	0.66 (0.51–0.79)	0.63 (0.52–0.73)	2.18 (1.32–3.60)
(ii) Youden *J* weighted for sensitivity	5	0.77 (0.65–0.86)	0.33 (0.22–0.45)	0.51 (0.40–0.61)	0.62 (0.45–0.77)	1.15 (0.93–1.41)
(iii) Youden *J* weighted for specificity	13	0.29 (0.19–0.42)	0.92 (0.83–0.97)	0.76 (0.55–0.91)	0.59 (0.50–0.68)	3.93 (1.36–11.34)
Females						
DSM-IV Abuse						
(i) Unweighted Youden *J*	5	0.72 (0.58–0.83)	0.56 (0.50–0.63)	0.28 (0.20–0.36)	0.90 (0.83–0.94)	1.64 (1.31–2.05)
(ii) Youden *J* weighted for sensitivity	2	0.92 (0.82–0.98)	0.20 (0.15–0.26)	0.21 (0.16–0.27)	0.92 (0.81–0.98)	1.16 (1.05–1.28)
(iii) Youden *J* weighted for specificity	10	0.25 (0.14–0.38)	0.86 (0.81–0.90)	0.29 (0.16–0.44)	0.83 (0.78–0.88)	1.76 (0.99–3.11)
DSM-IV dependence						
(i) Unweighted Youden *J*	7	0.71 (0.49–0.87)	0.70 (0.64–0.75)	0.18 (0.11–0.27)	0.96 (0.92–0.98)	2.34 (1.71–3.22)
(ii) Youden *J* weighted for sensitivity	2	1.00 (0.80–1.00)	0.19 (0.15–0.25)	0.10 (0.07–0.15)	1.00 (0.90–1.00)	1.24 (1.17–1.32)
(iii) Youden *J* weighted for specificity	11	0.46 (0.26–0.67)	0.90 (0.86–0.93)	0.30 (0.16–0.47)	0.95 (0.91–0.97)	4.55 (2.58–8.02)
DSM-5 disorders						
(i) Unweighted Youden *J*	6	0.63 (0.53–0.72)	0.74 (0.67–0.80)	0.59 (0.49–0.68)	0.77 (0.70–0.83)	2.42 (1.81–3.23)
(ii) Youden *J* weighted for sensitivity	2	0.96 (0.91–0.99)	0.26 (0.20–0.33)	0.44 (0.37–0.50)	0.92 (0.81–0.98)	1.30 (1.18–1.43)
(iii) Youden *J* weighted for specificity	11	0.23 (0.15–0.32)	0.93 (0.88–0.96)	0.65 (0.47–0.80)	0.67 (0.61–0.73)	3.11 (1.66–5.85)

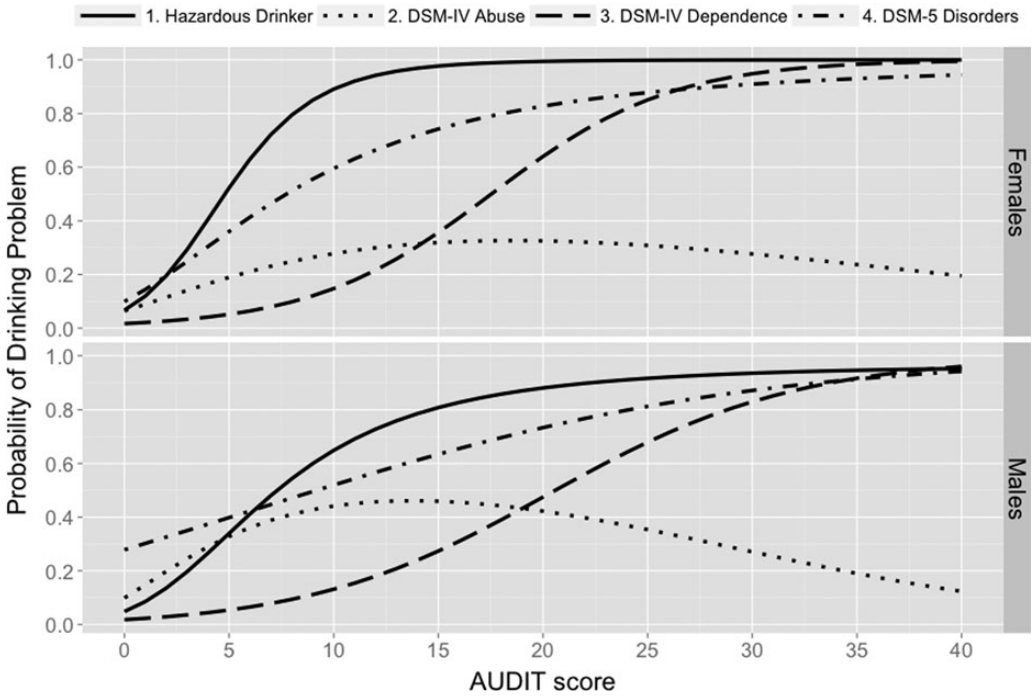

Fig. 1. Bayes' theorem post-test probability estimates for drinking problems in English males and females aged 18–35, according to AUDIT score.

If a male aged 18–35 has an AUDIT score of five, for example, this corresponds to a 0.34 probability of being a hazardous drinker, a 0.33 probability of being an alcohol abuser, a 0.05 probability of being alcohol dependent and a 0.40 probability of having an alcohol use disorder. Similarly, for a female, an AUDIT score of three would indicate a 0.29 probability of being a hazardous drinker, a 0.14 probability of being an alcohol abuser, a 0.03 probability of being alcohol dependent and a 0.25 probability of having an alcohol use disorder. For a male with an AUDIT score of 25, he would have a 0.92 probability of being a hazardous drinker, a 0.35 probability of being an alcohol abuser, a 0.68 probability of being alcohol dependent and a 0.81 probability of having an alcohol use disorder. For a female with an AUDIT score of 25, she would have a 1.00 probability of being a hazardous drinker, a 0.31 probability of being an alcohol abuser, a 0.85 probability of being alcohol dependent and a 0.88 probability of having an alcohol use disorder. It is worth noting that as AUDIT score increases, the likelihood of a positive alcohol abuse classification is replaced by a classification of dependence; this is because the a positive abuse classification is made only if the criteria for dependence are not met.

DISCUSSION

In this study we found that the AUDIT screening test was accurate for the assessment of reference standard classifications of hazardous drinking, DSM-IV alcohol abuse, DSM-IV alcohol dependence and DSM-5 alcohol use disorders in a sample of 18–35 year-old adults from UK primary care. The short-form AUDIT-C had a similar accuracy profile to the full AUDIT for the detection of hazardous drinking. The optimal cut-point threshold score varied substantially according to the choice of weighting for the Youden *J* index. Using Bayes' theorem, the post-test probability of drinking problems changed as AUDIT score increased, according to the slope of the probability curve.

If sensitivity is favoured over specificity, which may be more desirable when the costs of a false positive test are low, for example in a screening scenario which triggers further investigations or brief advice on healthy behaviour, then choosing a lower cut-point threshold score may be desirable from a clinical or population health perspective. However, individuals may value costs differently, and it is feasible that an individual would be upset or annoyed at being labelled as a heavy or risky drinker if in fact they are not; it also runs the theoretical risk of disengagement in false positive individuals, i.e. less likely to respond to health care advice/interventions. Alternatively, if the costs of a false positive are high, then specificity may be weighted more strongly than sensitivity. From a public health paradigm sensitivity is likely to be the dominant consideration when screening for hazardous alcohol consumption coupled with brief, low cost, interventions.

While using a single, simply applied threshold for screening and diagnostic tests may have advantages in practice settings, there are some disadvantages. Each point, or score, on a diagnostic scale or screening tool provides useful information that may be lost if scores are collapsed together into negative or positive categories. This simplistic approach implies that all test results above the threshold increase the likelihood that the condition or disease is present to exactly the same degree. However, if the likelihood associated with a range of different thresholds, for example each point on a screening test scale, can be calculated then more accurate estimates of the probability of a condition or disease can be made. This is particularly important when risk does not increase linearly with test score. This approach is also potentially useful in intervention research where AUDIT is used as an outcome measure: a change in AUDIT score will correspond to a change in probability, or

prevalence, of problem drinking (Foxcroft *et al.*, 2009). Whether thresholds or probability estimates are used, because AUDIT was designed as a screening rather than a diagnostic tool, either a positive threshold score or a high probability estimate for alcohol dependence or alcohol use disorders should trigger further tests, before any intervention.

It is possible that the non-random order of the presentation of the AUDIT and reference tests could have primed certain responses in the reference tests, though the 2-week delay between AUDIT and reference tests should protect against any priming bias. The order of presentation of the reference tests could also have led to a primed response for the TLFB, as TLFB always followed WMH-CIDI administration. Another potential limitation is that we used a 12-month time-frame for AUDIT responses, but the TLFB covered only the previous 90 days; so there is a mismatch in time-frame. However, we suggest that 90 days is a long enough period to be representative of the previous 12 months. Of those who agreed to participate in the study by returning the GLQ 41% were subsequently interviewed. But overall the questionnaire returns were low: only 2.2% of males and 3.4% of females who were approached both completed the AUDIT and participated in the telephone interview, resulting in a non-representative sample and a high risk of an external validity bias (Fernandez-Hermida *et al.*, 2012). In particular, the sample was skewed to low deprivation postcodes (i.e. the less deprived in the population), though there is an indication that drinking problem rates are broadly comparable with national sample surveys. In the 2007 Adult Psychiatric Morbidity Survey (APMS), alcohol dependence rates using the Severity of Alcohol Dependence Questionnaire (SADQ) for 18–35 year-olds are broadly comparable, given the different measurement tools used, to WMH-CIDI figures in the current study. For males, 16.8% were classified as alcohol dependent in APMS, and for females the figure was 4.7%. In the current study the figures were 13% males and 8.5% females. Also from APMS, binge drinking rates were 84% for men and 72% for women ('drunk six or more drinks on one occasion'), again broadly comparable given the different measurement approaches. In the current study the figures were 71% males and 66% females. There are no reasonably direct comparisons with APMS, or with the General Lifestyle Survey for England and Wales, to be made for the other measures and definitions we have used in the current study. We based our criteria for hazardous drinking on daily and weekly drinking levels defined by the UK Government (Royal College of Physicians, 2011). In sensitivity analyses, the results we report in this paper are robust to different definitions for hazardous drinking, though of course prevalence levels will vary according to definition.

Conclusions

Optimal test thresholds depended on the value attached to minimizing the cost associated with false test results. For screening tests it may be more appropriate to be more tolerant of false positives and use a threshold with higher sensitivity, though this is a matter for debate and further consideration by policy makers and clinicians.

An alternative approach, using Bayes' Theorem, is to calculate the post-test probability for each test score and to use this in feedback and dialogue with screened patients, which may include brief intervention or further tests. This approach has the advantage of using all available information rather than collapsing test scores above and below a selected threshold score. It will also take account of varying pre-test probabilities based on known prevalence rates for specific age, gender and other population parameters, and clinical judgment if desirable. Such an approach is entirely feasible using computer- or web-based assessment and feedback technology, and could also address some of the implementation problems that have been identified with alcohol screening and brief intervention in general practice (van Beurden *et al.*, 2012).

AUTHORS' CONTRIBUTION

D.R.F. and L.S. jointly conceived and designed the project and analysis. H.T. and S.H. collected data and with L.A.S. undertook initial descriptive analyses. D.R.F. undertook the analyses and led the writing for this paper. All authors contributed to reviewing the manuscript. D.R.F. and L.A.S. are guarantors for the paper.

Acknowledgements — We are grateful to the Thames Valley Primary Care Research Network for their help in recruitment of General Practices for the study.

Funding — This research was funded by Alcohol Research UK.

Conflict of interest statement. All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: all authors had financial support from Alcohol Research UK via a research grant to Oxford Brookes University for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous 3 years; no other relationships or activities that could appear to have influenced the submitted work.

REFERENCES

American Psychiatric Association. (1994) *Diagnostic and Statistical Manual of Mental Disorders (4th edn.) (DSM-IV)*. Washington, DC: APA.

American Psychiatric Association. (2013) *Diagnostic and Statistical Manual of Mental Disorders*, 5th edn. Arlington, VA: American Psychiatric Publishing.

Anderson P, Baumberg B. (2006) *Alcohol in Europe*. London: Institute of Alcohol Studies. http://ec.europa.eu/health/archive/ph_determinants/life_style/alcohol/documents/alcohol_europe_en.pdf (accessed 31 March 2014).

Babor TF, Biddle-Higgins JC, Saunders JB *et al.* (2001) *AUDIT: The Alcohol Use Disorders Identification Test: Guidelines for Use in Primary Health Care*. Geneva, Switzerland: World Health Organization.

Coulton S, Drummond C, James D *et al.* (2006) Opportunistic screening for alcohol use disorders in primary care: comparative study. *BMJ* **332**:511–7.

Fagan TJ. (1975) Nomogram for Bayes' theorem. *N Engl J Med* **293**:257.

Fernandez-Hermida JR, Calafat A, Becoña E *et al.* (2012) Assessment of generalizability, applicability and predictability (GAP) for evaluating external validity in studies of universal family-based prevention of alcohol misuse in young people: systematic methodological review of randomized controlled trials. *Addiction* **107**:1570–9.

Fischer JE, Bachmann LM, Jaeschke R. (2003) A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. *Intensive Care Med* **29**:1043–51.

Foxcroft DR, Kypri K, Simonite V. (2009) Bayes' theorem to estimate population prevalence from Alcohol Use Disorders Identification Test (AUDIT) scores. *Addiction* **104**:1132–7.

Goddard MJ, Hinbery I. (1990) Receiver operator characteristic (ROC) curves and non-normal data: an empirical study. *Stat Med* **9**:325–37.

Hall TW. (1985) Diagnosis, and Bayes' theorem. *Lancet* **325**: 705–6.
Katz MA. (1974) A probability graph describing the predictive value of a highly sensitive diagnostic test. *N Engl J Med* **291**:1115–6.
Kriston L, Holzel L, Weiser AK *et al*. (2008) Meta-analysis: are 3 questions enough to detect unhealthy alcohol use? *Ann Intern Med* **149**:879–88.
Li D, Shen F, Yin Y *et al*. (2013) Weighted Youden index and its two-independent-sample comparison based on weighted sensitivity and specificity. *Chin Med J* **126**:1150–54.
Lines C. (2010) *Appraisal for Screening for Alcohol Misuse: Report for the National Screening Committee*. Oxford: Solutions for Public Health http://www.screening.nhs.uk/policydb_download.php?doc=113 (accessed 2 May 2014).
McManus S, Meltzer H, Brugha T *et al*. (2009) Adult Psychiatric Morbidity in England, 2007: Results of a Household Survey: The NHS Information Centre for Health and Social Care.
R Development Core Team. (2008) R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.
Rehm J, Room R, van den Brink W *et al*. (2005) Alcohol use disorders in EU countries and Norway: an overview of the epidemiology. *Eur Neuropsychopharmacol* **15**:377–88.
Robin X, Turck N, Hainard A *et al*. (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**:77.
Robins LN, Wing J, Wittchen HU *et al*. (1988) The Composite International Diagnostic Interview: an epidemiologic instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Arch Gen Psychiatry* **45**:1069–77.
Royal College of Physicians. (2011) *Written Evidence to Science and Technology Select Committee Inquire on Alcohol Guidelines*. London: Royal College of Physicians. http://www.rcplondon.ac.uk/sites/default/files/rcp_evidence_to_the_inquiry_on_alcohol_guidelines_1.pdf (accessed 17 May 2014).
Simundic A-M. (2008) Measures of diagnostic accuracy: basic definitions. *EJIFCC* **19**. http://www.ifcc.org/ifcc-communications-publications-division-(cpd)/ifcc-publications/ejifcc-(journal)/e-journal-volumes/ejifcc-2008-vol-19/vol-19-no-4/measures-of-diagnostic-accuracy-basic-definitions/. (14 November 2014, date last accessed).
Smith LA, Foxcroft DR, Holloway A *et al*. (2014) Brief alcohol questionnaires for identifying hazardous, harmful and dependent alcohol use in primary care. *Cochrane Review*. In press.
Sobell LC, Sobell MB, Riley DM *et al*. (1988) The reliability of alcohol abusers' self-reports of drinking and life events that occurred in the distant past. *J Stud Alcohol* **49**:225–32.
Sobell LC, Sobell NB, Litten RZ *et al*. (1992) *Timeline Follow Back. A Technique for Assessing Self-Reported Alcohol Consumption Measuring Alcohol Use*. New Jersey: Humana Press Inc, 41–72.
Streiner DL, Cairney J. (2007) What's under the ROC? An introduction to receiver operating characteristic curves. *Can J Psychiatry* **52**:121–8.
van Beurden I, Anderson P, Akkermans RP *et al*. (2012) Involvement of general practitioners in managing alcohol problems: a randomized controlled trial of a tailored improvement programme. *Addiction* **107**:1601–11.
Wieand S, Gail MH, James BR *et al*. (1989) A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* **76**:585–92.
Youden WJ. (1950) An index for rating diagnostic tests. *Cancer* **3**:32–5.
Zweig MH, Campbell G. (1993) Receiver operator characteristic (ROC) plots; a fundamental evaluation tool in clinical medicine. *Clin Chem* **39**:561–77.