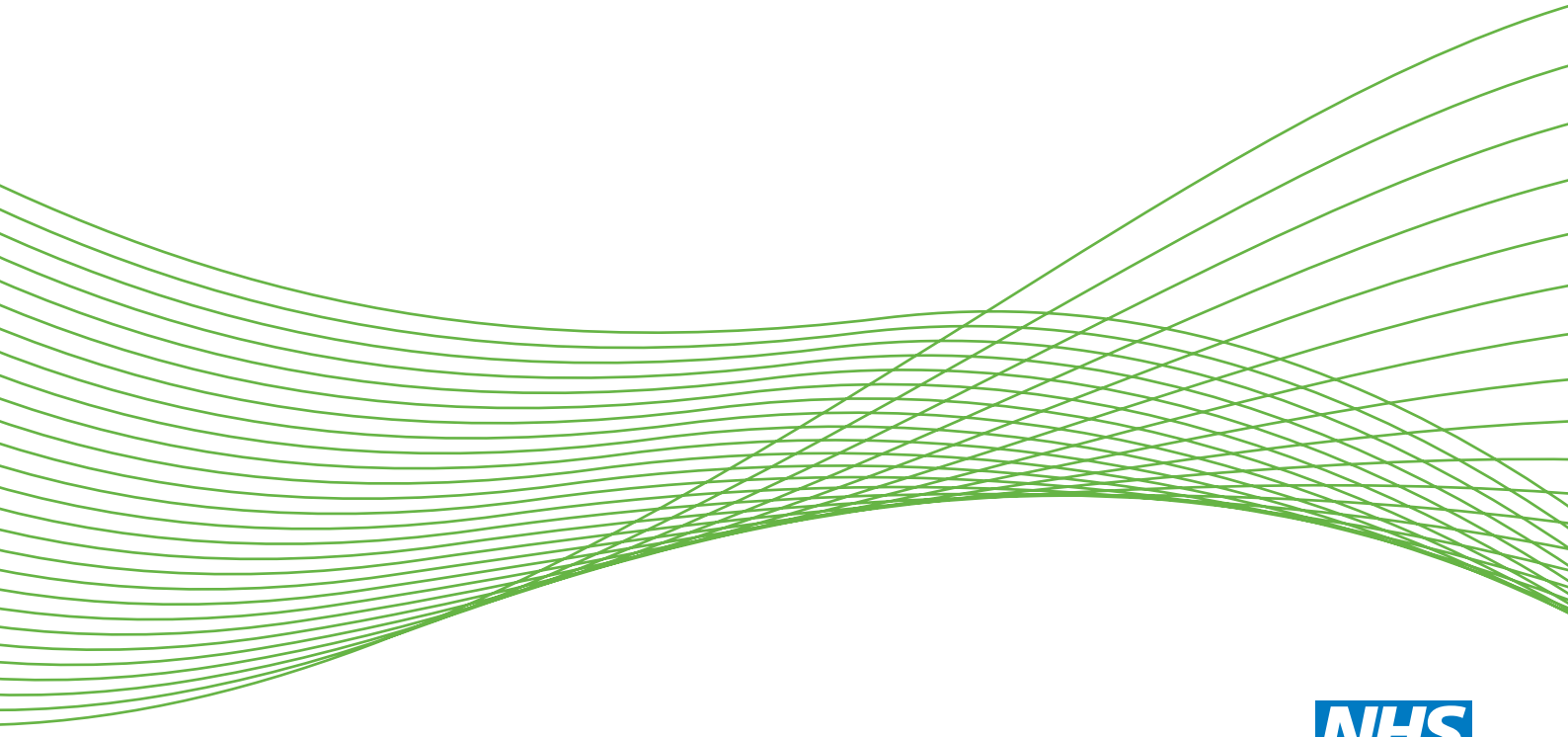
HEALTH TECHNOLOGY ASSESSMENT

VOLUME 19 ISSUE 1 JANUARY 2015
ISSN 1366-5278

Screening for psychological and mental health difficulties in young people who offend: a systematic review and decision model

Rachel Richardson, Dominic Trépel, Amanda Perry, Shehzad Ali, Steven Duffy, Rhian Gabe, Simon Gilbody, Julie Glanville, Catherine Hewitt, Laura Manea, Stephen Palmer, Barry Wright and Dean McMillan



NHS
National Institute for Health Research

DOI 10.3310/hta19010

Screening for psychological and mental health difficulties in young people who offend: a systematic review and decision model

Rachel Richardson,[1] Dominic Trépel,[1] Amanda Perry,[1] Shehzad Ali,[1] Steven Duffy,[2] Rhian Gabe,[1,3] Simon Gilbody,[1,3] Julie Glanville,[2] Catherine Hewitt,[1] Laura Manea,[1,3] Stephen Palmer,[4] Barry Wright[1,3] and Dean McMillan[1,3]*

[1]Department of Health Sciences, University of York, York, UK
[2]York Health Economics Consortium, York, UK
[3]Hull York Medical School, University of York, York, UK
[4]Centre for Health Economics, University of York, York, UK

*Corresponding author

Declared competing interests of authors: Simon Gilbody is a member of the National Institute for Health Research Health Technology Assessment Clinical Evaluation and Trials Board.

Published January 2015
DOI: 10.3310/hta19010

This report should be referenced as follows:

Richardson R, Trépel D, Perry A, Ali S, Duffy S, Gabe R, *et al*. Screening for psychological and mental health difficulties in young people who offend: a systematic review and decision model. *Health Technol Assess* 2015;**19**(1).

Health Technology Assessment is indexed and abstracted in *Index Medicus*/MEDLINE, *Excerpta Medica*/EMBASE, *Science Citation Index Expanded* (SciSearch®) and *Current Contents®*/ Clinical Medicine.

Health Technology Assessment

HTA/HTA TAR

ISSN 1366-5278 (Print)

ISSN 2046-4924 (Online)

Impact factor: 5.116

Health Technology Assessment is indexed in MEDLINE, CINAHL, EMBASE, The Cochrane Library and the ISI Science Citation Index and is assessed for inclusion in the Database of Abstracts of Reviews of Effects.

This journal is a member of and subscribes to the principles of the Committee on Publication Ethics (COPE) (www.publicationethics.org/).

Editorial contact: nihredit@southampton.ac.uk

The full HTA archive is freely available to view online at www.journalslibrary.nihr.ac.uk/hta. Print-on-demand copies can be purchased from the report pages of the NIHR Journals Library website: www.journalslibrary.nihr.ac.uk

Criteria for inclusion in the *Health Technology Assessment* journal

Reports are published in *Health Technology Assessment* (HTA) if (1) they have resulted from work for the HTA programme, and (2) they are of a sufficiently high scientific quality as assessed by the reviewers and editors.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

HTA programme

The HTA programme, part of the National Institute for Health Research (NIHR), was set up in 1993. It produces high-quality research information on the effectiveness, costs and broader impact of health technologies for those who use, manage and provide care in the NHS. 'Health technologies' are broadly defined as all interventions used to promote health, prevent and treat disease, and improve rehabilitation and long-term care.

The journal is indexed in NHS Evidence via its abstracts included in MEDLINE and its Technology Assessment Reports inform National Institute for Health and Care Excellence (NICE) guidance. HTA research is also an important source of evidence for National Screening Committee (NSC) policy decisions.

For more information about the HTA programme please visit the website: http://www.nets.nihr.ac.uk/programmes/hta

This report

The research reported in this issue of the journal was funded by the HTA programme as project number 10/35/01. The contractual start date was in August 2011. The draft report began editorial review in May 2013 and was accepted for publication in January 2014. The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The HTA editors and publisher have tried to ensure the accuracy of the authors' report and would like to thank the reviewers for their constructive comments on the draft document. However, they do not accept liability for damages or losses arising from material published in this report.

Published by the NIHR Journals Library (www.journalslibrary.nihr.ac.uk), produced by Prepress Projects Ltd, Perth, Scotland (www.prepress-projects.co.uk).

Abstract

Screening for psychological and mental health difficulties in young people who offend: a systematic review and decision model

Rachel Richardson,[1] Dominic Trépel,[1] Amanda Perry,[1] Shehzad Ali,[1] Steven Duffy,[2] Rhian Gabe,[1,3] Simon Gilbody,[1,3] Julie Glanville,[2] Catherine Hewitt,[1] Laura Manea,[1,3] Stephen Palmer,[4] Barry Wright[1,3] and Dean McMillan[1,3]*

[1]Department of Health Sciences, University of York, York, UK
[2]York Health Economics Consortium, York, UK
[3]Hull York Medical School, University of York, York, UK
[4]Centre for Health Economics, University of York, York, UK

*Corresponding author dean.mcmillan@york.ac.uk

Background: There is policy interest in the screening and treatment of mental health problems in young people who offend, but the value of such screening is not yet known.

Objectives: To assess the diagnostic test accuracy of screening measures for mental health problems in young people who offend; to evaluate the clinical effectiveness and cost-effectiveness of screening and treatment; to model estimates of cost; to assess the evidence base for screening against UK National Screening Committee criteria; and to identify future research priorities.

Data sources: In total, 25 electronic databases including MEDLINE, PsycINFO, EMBASE and The Cochrane Library were searched from inception until April 2011. Reverse citation searches of included studies were undertaken and reference list of included studies were examined.

Review methods: Two reviewers independently examined titles and abstracts and extracted data from included studies using a standardised form. The inclusion criteria for the review were (1) population – young offenders (aged 10–21 years); (2) intervention/instrument – screening instruments for mental health problems, implementation of a screening programme or a psychological or pharmacological intervention as part of a clinical trial; (3) comparator – for diagnostic test accuracy studies, any standardised diagnostic interview; for trials, any comparator; (4) outcomes – details of diagnostic test accuracy, mental health outcomes over the short or longer term or measurement of cost data; and (5) study design – for diagnostic test accuracy studies, any design; for screening programmes, randomised controlled trials or controlled trials; for clinical effectiveness studies, randomised controlled trials; for economic studies, economic evaluations of screening strategies or interventions.

Results: Of 13,580 studies identified, nine, including eight independent samples, met the inclusion criteria for the diagnostic test accuracy and validity of screening measures review. Screening accuracy was typically modest. No studies examined the clinical effectiveness of screening, although 10 studies were identified that examined the clinical effectiveness of interventions for mental health problems. There were too few studies to make firm conclusions about the clinical effectiveness of treatments in this population. No studies met the inclusion criteria for the assessment of the cost-effectiveness of screening or treatment. An exemplar decision model was developed for depression, which identified a number of the likely key

drivers of uncertainty, including the prevalence of unidentified mental health problems, the severity of mental health problems and their relationship to generic measures of outcome and the impact of treatment on recidivism. The information evaluated as part of the review was relevant to five of the UK National Screening Committee criteria. On the basis of the above results, none of the five criteria was met.

Limitations: The conclusions of the review are based on limited evidence. Conclusions are tentative and the decision model should be treated as an exemplar.

Conclusions: Evidence on the clinical effectiveness and cost-effectiveness of screening for mental health problems in young people who offend is currently lacking. Future research should consider feasibility trials of clinical interventions to establish important parameters ahead of conducting definitive trials. Future diagnostic studies should compare the diagnostic test accuracy of a range of screening instruments, including those recommended for use in the UK in this population. These studies should be designed to reduce the decision uncertainty identified by the exemplar decision model.

Registration: This study is registered as PROSPERO CRD42011001466.

Funding: The National Institute for Health Research Health Technology Assessment programme.

Contents

List of tables

List of figures

List of boxes

List of abbreviations

ADH	attention deficit hyperactivity
ADHD	attention deficit hyperactivity disorder
ANOVA	analysis of variance
BDI	Beck Depression Inventory
CBT	cognitive–behavioural therapy
CHAT	Comprehensive Health Assessment Tool
CI	confidence interval
CWD-A	Adolescent Coping with Depression
DFD	depression-free day
DISC	Diagnostic Interview Schedule for Children
DOR	diagnostic odds ratio
DPS	Diagnostic Interview Schedule for Children Predictive Scales
DSM	*Diagnostic and Statistical Manual of Mental Disorders*
EVPI	expected value of perfect information
HRSD	Hamilton Rating Scale for Depression
ICER	incremental cost-effectiveness ratio
IES	Impact of Events Scale
K-SADS	Schedule for Affective Disorders and Schizophrenia for School-Age Children
MAS	Manifest Anxiety Scale
MAYSI-2	Massachusetts Youth Screening Instrument – version 2
MFQ	Mood and Feelings Questionnaire
MMPI-A	Minnesota Multiphasic Personality Inventory – Adolescent version
Nacro	National Association for the Care and Resettlement of Offenders
NHS EED	NHS Economic Evaluation Database
NICE	National Institute for Health and Care Excellence
NSC	National Screening Committee
ODD	oppositional defiant disorder
PICO	population/patient problem, intervention, comparison, outcome
PSS-SR	Post-Traumatic Stress Disorder Symptom Scale Self Report
PTSD	post-traumatic stress disorder
QALY	quality-adjusted life-year
QUADAS-2	Quality Assessment Tool for Diagnostic Accuracy Studies – version 2
RDC	Research Diagnostic Criteria
SADS	Schedule for Affective Disorders and Schizophrenia
SADS-L	Schedule for Affective Disorders and Schizophrenia – Lifetime version
SD	standard deviation
SIfA	Screening Interview for Adolescents
SMD	standardised mean difference
SNASA	Salford Needs Assessment Scale for Adolescents
UCLA PTSD RI	University of California at Los Angeles Post-Traumatic Stress Disorder Reaction Index
VoI	value of information
WTP	willingness to pay
YSR	Youth Self Report scale

Plain English summary

Young people who have offended are more likely than people who have not offended to have mental health problems and they are also more likely to offend again. It may, therefore, be important to identify the mental health difficulties in this group and give them help for these problems.

There are, however, a number of unanswered questions about identifying mental health problems in young people who offend. These include:

- How accurate are the different ways of identifying these difficulties?
- If a difficulty is identified, how well does any treatment given for this difficulty work?
- Does identifying mental health problems in this way represent good value for money?

We sought to identify all research that could help to answer these questions. We identified a small number of studies that looked at how accurate different tools were at identifying mental health problems in this group. Most tools had limited accuracy. We also identified a small number of studies that had looked at whether or not treatments work for mental health difficulties in young people who offend. Although there was some encouraging evidence, it remains uncertain if treatments are effective in this group. In general, our search identified few studies and those studies we did identify were often of low quality.

There is a need for future studies that establish how effective and cost-effective treatments are for these difficulties. There is also a need for future studies that better establish how accurate screening instruments are for identifying mental health problems.

Scientific summary

Background

Young people who offend are at an increased risk of a range of mental health problems including depression, anxiety and disruptive disorders, including conduct disorder and attention hyperactivity deficit disorder (ADHD). These mental health difficulties are associated with a number of negative consequences both for the young person and for society, such as an increased risk of reoffending. Despite this, mental health problems remain underdetected and undertreated in young people who offend. In recognition of this, there is currently policy interest in screening for mental health problems in this population. Although mental health screening is currently recommended for young people who offend, the value of this is currently unknown.

Objectives

The review had five objectives:

1. to conduct a systematic review and evidence synthesis of the diagnostic properties and validity of existing screening measures for mental health problems in young people who offend
2. to assess the clinical effectiveness of screening strategies in this population and (more broadly) to assess the clinical effectiveness of interventions for mental health problems
3. to assess the cost-effectiveness of screening strategies in this population and (more broadly) to assess the cost-effectiveness of interventions for mental health problems, with specific reference to identifying in which groups they may be cost-effective
4. to assess whether or not current screening strategies meet minimum criteria laid down by the UK National Screening Committee (NSC) in the light of this evidence synthesis
5. to identify research priorities and the value of developing future research into screening strategies for young offenders with mental health problems.

Methods

A single, comprehensive search of the literature was undertaken to identify literature relevant to each stage of the review. In total, 25 electronic databases were searched, including MEDLINE, PsycINFO, EMBASE and The Cochrane Library. Each database was searched from inception until April 2011. Internet resources of relevant organisations and conference proceedings were also examined. Sources of data spanned the health, mental health and criminal justice literature.

Reverse citation searches of included studies were undertaken and reference list of included studies and previous reviews were also examined. Experts in the field were contacted to identify other potentially relevant literature.

After deduplication, 13,580 studies were examined for potential inclusion, of which 219 were selected for further evaluation. Data were extracted to a standardised coding sheet for all studies meeting the inclusion criteria. At each stage, two reviewers independently examined citations and extracted data. Disagreements were resolved by consensus or deferred to a third party if necessary.

Separate inclusion and exclusion criteria were developed for each phase of the review; these can be broadly summarised as follows:

- *population:* young people (aged 10–21 years) who have offended and who are in contact with the criminal justice system
- *intervention/instrument:* screening instruments for mental health problems, implementation of a screening programme or psychological or pharmacological interventions as part of a clinical trial
- *comparator:* for diagnostic test accuracy studies, a standardised diagnostic interview conducted to internationally recognised standards; for screening programmes, any comparator
- *outcomes:* details of diagnostic test accuracy, mental health outcomes over the short or longer term or any measurement of cost data
- *study design:* for diagnostic test accuracy studies, any design; for screening programmes, randomised controlled trials or controlled trials; for clinical effectiveness studies, randomised controlled trials; and for economic studies, economic evaluations of screening strategies or interventions.

Evidence was sought across a range of mental health difficulties in young people who offend, including depressive disorders, anxiety disorders, disruptive disorders and other disorders such as psychosis and autistic spectrum disorders, and self-harm and suicidal behaviour. There were too few studies to conduct a meta-analysis for any stage of the review and so a series of narrative syntheses was undertaken.

To evaluate the cost-effectiveness of identification strategies, the policy question addressed by the decision model was constrained to focus on the screening and subsequent management of one common mental health problem in the young offender population: depression. The rationale for constraining the policy question and developing an 'exemplar' case study for the decision model was that (1) depression is highly prevalent in young offenders; (2) taken together there is more evidence on screening and treatment effectiveness for depression in young offenders than for other mental health conditions; (3) depression-related health states could be mapped onto health-related quality of life (or utility) measures [e.g. quality-adjusted life-years (QALYs)]; and (4) depression is not an externalising condition and may, therefore, go undetected.

Results

Nine studies including eight independent samples met the inclusion criteria for the diagnostic test accuracy and validity of screening measures review. The Massachusetts Youth Screening Instrument – version 2 (MAYSI-2) was the most commonly used screening measure. Data for the MAYSI-2 suggested moderate sensitivity and specificity at standard cut-off points commonly cited in the literature. Firm conclusions could not be made because of the low number of included studies for any one combination of mental health problem and screening measure. However, data were identified on screening accuracy for some mental health problems, including depression, post-traumatic stress disorder (PTSD), ADHD, conduct disorder and oppositional defiant disorder (ODD). There appeared to be no evidence that screening measures specifically designed for use in young offender groups such as the MAYSI-2 had superior operating characteristics to more general measures.

No studies were identified that examined the clinical effectiveness of screening. Ten studies met the inclusion criteria for the examination of clinical effectiveness. Of the included studies, some interventions targeted depression, anxiety including PTSD, conduct disorder, ODD and ADHD, while other interventions had a broader focus (e.g. improving interpersonal functioning). There were too few studies for any one combination of intervention and outcome to make firm conclusions about the clinical effectiveness of treatments for mental health problems, particularly because the quality assessment indicated either an unclear or a high risk of bias for many of the studies.

No studies met the inclusion criteria for the assessment of the cost-effectiveness of screening or treatment.

On the basis of the data identified in the systematic reviews, an exemplar decision model for depression was developed to provide initial insights into the possible merits of identification and treatment strategies and the importance of perspectives adopted given the intersectoral nature of this question. However, these insights need to be considered within the limitations of the available evidence emerging from the systematic review of diagnostic and clinical effectiveness studies. Nonetheless, the decision model makes a contribution to the overall evidence by providing an exemplar based on a formal quantitative framework that provides a clear indication of the various inputs and data sources required to appropriately inform cost-effectiveness assessment. Although formal value of information analysis was not feasible, deterministic sensitivity analysis highlighted key drivers of the model, which should inform future research design. These include identifying the level of previously undetected mental health problems in this population, the importance of using generic measures to permit the calculation of QALYs, and assessing the impact of mental health treatment on intersectoral outcomes, including recidivism. Importantly, the model provides an iterative basis for updating and revisiting the findings as new evidence emerges in the future.

The results of the evidence synthesis were used to assess whether or not UK NSC criteria were met for screening for mental health problems in young people who offend. Five of the UK NSC criteria could be examined on the basis of the current review; these included the existence of a precise and valid screening instrument (UK NSC criterion 5), a known distribution of test values and a cut-off agreed for the instrument (criterion 6), the existence of an effective treatment (criterion 10), evidence from randomised controlled trials that screening is effective (criterion 13) and opportunity costs should be economically balanced in relation to expenditure (criterion 16). None of the criteria was met on the basis of the evidence examined as part of this review.

Conclusions

Screening is only of value if there is an effective intervention, and this has not yet been established for the treatment of mental health problems in this population. In terms of clinical effectiveness, the limitations of the existing randomised controlled trial evidence base suggest that further feasibility trials of clinical effectiveness are needed to establish important parameters ahead of definitive trials of effectiveness in this area. As indicated by the decision model, future trials should gather information to permit the calculation of QALYs and should seek to assess whether or not treatment alters intersectoral outcomes, particularly recidivism.

Future research priorities for diagnostic test accuracy include validation studies in which the performance of a range of screening measures is directly compared against a gold standard diagnostic interview conducted to internationally recognised criteria. Screening measures currently recommended for use in the UK to identify mental health difficulties among young people who have offended, specifically the mental health screen of the Comprehensive Health Assessment Tool, should be directly compared with other available screening measures as part of such studies. As indicated by the decision model, studies should seek to calculate the diagnostic performance of measures in identifying previously unidentified cases. This fundamental work on clinical effectiveness and diagnostic test accuracy should be conducted ahead of a trial of screening in this area. Evidence was lacking for both community and incarcerated settings, so these recommendations apply equally to both settings.

Study registration

This study is registered as CRD42011001466.

Funding

Funding for this study was provided by the Health Technology Assessment programme of the National Institute for Health Research.

Chapter 1 Introduction and background

Mental health difficulties in young people who offend

In England and Wales young people between the ages of 10 and 17 years committed 201,800 offences in 2009–10 and were responsible for 17% of all proven offending.[1] Problems linked to offending behaviour include educational underachievement, substance abuse and mental illness. With regard to mental illness, there is a lack of precise estimates of the prevalence and types of mental health difficulties experienced by young people who offend, but what evidence there is suggests prevalence figures substantially in excess of age-equivalent, general population rates.[2,3]

The types of difficulties for which these rates are elevated cover a wide range of mental health problems, including depression, anxiety [particularly post-traumatic stress disorder (PTSD)], attention deficit hyperactivity disorder (ADHD), psychotic-like symptoms and self-harm.[3,4] There is also some evidence that rates of learning disability and special educational needs may be high among this group.[5] The presence of mental health difficulties among young people who offend may increase the risk of a range of negative outcomes for both the young person and the wider community. The presence of mental health problems such as depression among this group may act as a risk factor for the persistence of offending behaviour into adulthood.[6] Additionally, conduct disorder in young people leads to a range of difficulties. For example, adults who had conduct disorder in adolescence are 70 times more likely to be imprisoned before the age of 25 years.[7] The cost of crime in England and Wales committed by adults who had conduct disorder as a child has been estimated at £2.25B.[8]

Screening for mental health difficulties in young people who offend

Despite their prevalence and potential to increase the risk of negative outcomes, these difficulties remain under-identified and under-treated.[4,9] A number of screening methods have been used to identify young people with mental health problems, including those specifically tailored to young people who offend [e.g. Massachusetts Youth Screening Instrument – version 2 (MAYSI-2)[10]] and those developed in general settings.

Interventions for mental health difficulties in young people who offend

Screening, however, can be justified only if it results in a more effective treatment than would otherwise be the case and does so with a favourable ratio of costs to benefits.[11] There is substantial evidence from studies of young non-offender samples that effective psychological and pharmacological treatments exist for many of the mental health problems that are common in young people who offend, including treatments for depression,[12] anxiety problems,[13–15] ADHD[16] and psychotic-like symptoms.[17] Evidence on the effectiveness of mental health interventions specifically for young people who offend appears to be very limited. One previous systematic review in this area cautiously concluded that treatments may be effective, although it suggested that larger, high-quality trials were needed.[18]

Current UK policy and practice

In the UK, all young people who have offended are the responsibility of youth offending teams. The majority of young people are supervised in the community, but a smaller proportion is given a sentence that includes a custodial component, which is then followed by subsequent supervision in the community. A number of factors determine whether a young person is given a custodial sentence or is supervised in the community; these include the severity of the crime and the extent of any previous offending behaviour.

Those young people supervised in the community will typically receive either a referral order, if a first offence, or a youth rehabilitation order. These orders can vary in terms of both the level of supervision required and the additional conditions placed on the young person. For those people given a custodial sentence, the setting in which the person is placed will be determined by age and level of maturity. Secure children's homes are typically for young people aged 10–15 years and young offender institutions are for those aged ≥ 16 years. There is also the option of a secure training centre for those aged 15 or 16 years.

There has been a focus in the UK in recent years on the prevention of offending by young people and the treatment of these young people.[19] As a result, there is substantial policy interest in screening for mental health problems in people who offend and access to appropriate mental health services for such people. The Criminal Justice Bill 2007 and Lord Bradley's report[20] set out a range of strategies to improve the situation, including youth rehabilitation orders. Department of Health guidelines[21] have also made it clear that young people in secure settings should have appropriate access to mental health services. Such policy documents are supported by the recent national framework to improve mental health and well-being,[22] which organises six high-level objectives of mental health strategy across all sectors of society.

Available evidence on practice suggests that the provision of adequate mental health screening and intervention remains patchy, with demand outstripping supply.[4] A joint initiative between the Youth Justice Board and the Department of Health has sought to improve the identification and assessment of health-related needs in children and young people in contact with the youth justice system, including mental health needs. This initiative has led to the Comprehensive Health Assessment Tool (CHAT),[23] a bespoke, strategic toolkit that aims to identify and assess health-related needs of children and young people in contact with any part of the youth justice system. Within CHAT there are five areas of assessment:

- part 1 assesses any immediate risk associated with physical health, mental health, substance misuse and safety
- part 2 assesses physical health
- part 3 assesses substance misuse
- part 4 assesses mental health
- part 5 assesses neurodevelopment disorders such as learning disability, autistic spectrum disorders and speech and language impairment, as well as any traumatic brain injury.

Versions of CHAT are available for the secure estate and community settings. For the secure estate, a reception health screen needs to be completed within 2 hours or before the first night of admission to identify any immediate risks or concerns, which may lead to the fast-tracking of a more detailed CHAT evaluation for any areas identified as important. All areas of the CHAT assessment should then be completed in the first 10 days of intake, with the mental health assessment being completed within the first 3 days. For community settings the health reception screen is not used.

It is intended that the information from an updated version of Asset,[24] termed AssetPlus, will be made available to the professionals conducting the CHAT assessment in both incarcerated and community settings. Asset is an assessment tool that aims to identify those risk factors for the young person's offending, including mental health difficulties.[24] It can be used, therefore, to identify potential mental health needs that may

require further assessment and intervention. CHAT will replace previous mental health screening pathways, which, following a red flag on the Asset tool, involved further structured assessment with measures such as the Screening Interview for Adolescents (SIfA).[25]

The Comprehensive Health Assessment Tool began to be rolled out in 2012. The clinical effectiveness and cost-effectiveness of this screening strategy has not yet been determined.

Summary

A wide range of mental health problems are common in young people who offend, and their presence is linked to a range of negative consequences both for the young person and for the wider society. There is currently substantial policy interest in screening for mental health problems in young people who offend, but the value of such screening is currently unknown.

Chapter 2 Description of the decision problem

The purpose of this research was to apply rigorous systematic review and evidence synthesis techniques to answer the question, 'What would be the benefits of carrying out a screening assessment for treatable psychological and mental health conditions in young offenders and in which groups might it be cost-effective?'

Current UK policy provides guidance on screening for mental health problems in young people who offend,[23] as described in the previous chapter, but the clinical effectiveness and cost-effectiveness of the recommended screening pathways is largely unknown. There are, in fact, a number of ways in which screening pathways could be configured and a large number of uncertainties exist. The decision problem can be framed in terms of these uncertainties. A main aim of the review is to establish the extent to which existing evidence can reduce these uncertainties and to identify where future research should be targeted so that uncertainties can be further reduced.

Screening

One option for identifying mental health problems in young people who offend would be to offer this entire group a detailed diagnostic mental health assessment in the form of a gold standard interview conducted to internationally recognised criteria.[26,27] There are advantages to this: all who were offered treatment would be in need of it and all of those not given treatment would not require it. Although such an approach would give perfect precision, it may not be feasible because it may require substantial resources to implement.

The use of screening instruments, which trade a saving in resources for a reduction in precision, is the typical alternative to such a strategy.[11] Screening measures that have been used with young offenders can be divided into a number of broad categories: those that are designed to detect a specific mental health problem, such as major depression, and those that are designed to detect a general mental health problem or need.[28] Often this maps onto a division in young offender measures between those instruments that provide diagnostic test accuracy data and those that identify a mental health need but do not establish the accuracy against a gold standard diagnostic interview.

A further division is into those measures that are specifically designed for use with a young offender population (e.g. MAYSI-2,[10] CHAT mental health screen[23]) and those that are used with young offenders but which were originally developed for use in the wider population. A potential advantage of measures designed specifically for young offenders is that they may consider expected characteristics of the population (e.g. limited literacy) and may be designed for use by youth justice personnel with no formal mental health training. However, a potential disadvantage is that they may not have received the same level of psychometric evaluation as some of the more widely used measures.

Each screening instrument from these broad categories could be used in a number of ways to make a decision about a person's mental health needs, including the need for treatment. Scores on a screening measure could be considered alone in making that decision, in combination with each other (e.g. a general screen for any mental health problem followed by a disorder-specific screen) or in combination with a gold standard (e.g. a general screen followed by gold standard interview for all those scoring positively on the screen).

Currently, there is uncertainty around which broad category of instrument is likely to be most effective (e.g. bespoke measures for young offenders vs. measures originally designed for use in the wider population) and within a category it is unclear if particular screening instruments are more accurate than others in identifying mental health problems. In addition, there is further uncertainty around whether a decision should be made on the basis of a single instrument or whether a combination should be used in a screening pathway.

Many screening instruments have a range of possible scores and so it is possible to identify different points along that range above which a person could be predicted by the screening instrument to have a mental health difficulty. As this cut-off point is varied, sensitivity and specificity will also change in a consistent way: as sensitivity increases, specificity will decrease (and vice versa).[11] (For an introduction to methods of quantifying diagnostic test accuracy, including concepts such as sensitivity and specificity, see *Appendix 1*.) There is, then, always a balance to be struck: if sensitivity is high, specificity is likely to be low; if specificity is high, sensitivity is likely to be low. A decision needs to be made about what balance between sensitivity and specificity is likely to be appropriate in a particular decision context. There are no definitive guidelines but, as a general rule, when the clinical context involves screening, high sensitivity is usually valued over high specificity. If sensitivity is high, this means that few people who have a condition will be missed, even if this is at the expense of somewhat lower specificity. Ensuring that few people with the condition are missed is often an aim of a screening strategy. However, in many decision contexts – including screening for mental health problems in young people who offend – it may not be possible to ensure very high sensitivity. Screening measures for mental health problems can have substantial inaccuracies when assessed against a gold standard, which means that very high sensitivity on such instruments is likely to be associated with low specificity. A consequence of low specificity is a high false-positive rate, which can be problematic in a number of ways. For example, if screening is used in the absence of a confirmatory gold standard diagnosis, treatment may be offered to many people who do not in fact require it. This may be potentially damaging to the recipients and can have substantial costs attached to it for services. Even if a screening measure is followed by a confirmatory diagnostic assessment, it may be inefficient and prohibitively costly to refer on for that further assessment all people who score positive to a screen if that number contains a large number of false positives. As a very broad guideline, then, a cut-off on a screening instrument may be required that gives sufficiently high sensitivity while retaining moderate specificity.

Studies of diagnostic test accuracy typically evaluate the screening measure against a gold standard categorisation of those with and without the mental health diagnosis, regardless of whether or not the true cases are already known to services or are previously unidentified cases. In this particular decision context, the screening for mental health problems in young people who offend, screening may be of value only for the identification of previously unidentified cases, because known cases may already be receiving treatment. There are a number of uncertainties related to this distinction between known and unidentified cases. It is unclear if the diagnostic performance of the test may differ if restricted to the identification of previously unknown cases. It is also unclear if the characteristics of the previously unidentified cases and the already identified cases differ, and this may be of relevance to understanding the likely performance characteristics of a test when restricted to the identification of new cases. For example, it is possible that already known cases will be more severe and therefore easily identifiable in the absence of screening, whereas unidentified cases may be less severe. This may have consequences for the need to offer treatment or the type of treatment offered. The prevalence of unidentified cases is also unclear, and this may have consequences for the balance between true positives and false positives at a particular cut-off point on an instrument. This in turn may affect the selection of an optimal cut-off point and the balance it offers between sensitivity and specificity.

Additional features of the decision problem relate to uncertainties about the behaviour of professionals in terms of screening. For example, it is unknown whether or not professionals find particular instruments acceptable and whether or not the results from a screening measure have an impact on professionals' behaviour, such as making a referral for a particular type of treatment.

Clinical effectiveness and cost-effectiveness

On the assumption that professional behaviour is altered by the results of a screening test, screening and referring are of use only if there is an effective and cost-effective treatment for the particular mental health problem. In terms of effectiveness there are a large number of uncertainties. These include whether or not interventions for mental health problems in young people who offend are clinically effective and cost-effective, whether or not improvements in mental health symptoms are related to changes in other outcomes, such as the likelihood of reoffending, whether or not the interventions are acceptable to this population and whether or not potentially effective interventions can be feasibly delivered in UK settings.

Setting

Young people who have offended may be in the community or incarcerated. In terms of the decision problem outlined above, each of the considerations applies separately to these two settings. It is possible, for example, that a distinct screening pathway may be more appropriate in one setting than in another.

Objectives

On the basis of this decision problem we developed five objectives related to diagnostic test accuracy, the clinical effectiveness and cost-effectiveness of screening and (more broadly) the clinical effectiveness and cost-effectiveness of interventions for mental health problems in young people who offend. These five objectives are to:

1. conduct a systematic review and evidence synthesis of the diagnostic properties and validity of existing screening measures for mental health problems in young people who offend
2. assess the clinical effectiveness of screening strategies in this population and (more broadly) the clinical effectiveness of interventions for mental health problems
3. assess the cost-effectiveness of screening strategies in this population and (more broadly) the cost-effectiveness of interventions for mental health problems, with specific reference to identifying in which groups they may be cost-effective
4. assess whether or not current screening strategies meet minimum criteria laid down by the UK National Screening Committee (NSC) in the light of this evidence synthesis
5. identify research priorities and the value of developing future research into screening strategies for young offenders with mental health problems.

Structure of the report

We carried out a single comprehensive search to identify the evidence needed for this research. This search is described in *Chapter 3*. We then conducted the research in a number of interlinked phases in which we summarised the available literature on screening assessments for treatable psychological and mental health conditions in young offenders.

At each stage of the review and in the production of the final report we adhered to the relevant guidelines for the conduct and reporting of systematic reviews.[29,30] The research is registered on the PROSPERO database (registration number CRD42011001466). A copy of the original protocol for the review is available alongside copies of this report on the National Institute for Health Research (NIHR) website (www.journalslibrary.nihr.ac.uk/).

Stakeholder involvement

We established an expert advisory group and two stakeholder groups. The expert advisory group consisted of academics with methodological expertise in the conduct of systematic reviews and content expertise in the criminal justice system. Members of this group were approached at various stages of the project to offer advice on specific questions.

One stakeholder group consisted of professionals working within the justice system. We sought to include professionals working in both community settings and the secure estate. We met with members of this stakeholder group at various stages of the project. A specific role of this group was to help establish current UK practice in the screening and treatment of mental health problems in young people who offend and more generally to clarify the nature of the decision problem.

A second stakeholder group consisted of young people (age range 10–15 years) from the National Association for the Care and Resettlement of Offenders (Nacro). We held two meetings with these young people to gather their views on a range of subjects relevant to the review, including the acceptability of different potential screening pathways and different types of interventions. The older members of this group were asked to comment and help draft the plain English summary.

Appendix 2 provides a list of the stakeholders and professionals who provided advice during the review process.

Chapter 3 Literature search

Literature searches were undertaken to identify studies about the screening, clinical effectiveness and cost-effectiveness of psychological and mental health difficulties in young people who offend.

Search terms

The search strategies were devised using a combination of subject indexing terms, such as medical subject heading (MeSH) in MEDLINE, and free-text search terms in the title and abstract. The search terms were identified through discussion among the research team, through contact with members of the advisory group, by scanning background literature and by browsing database thesauri.

We considered two main approaches to searching the literature: a single comprehensive search to identify studies of relevance to each phase of the review (e.g. screening, clinical effectiveness, cost-effectiveness) and an alternative strategy of developing separate searches for each phase. We chose to use a single comprehensive search as the most effective and efficient means of identifying the relevant literature for each phase.

The search terms for each database covered three broad constructs:

- age: terms to identify adolescents or young people
- offenders: terms to identify people who had offended or who were in contact with the criminal justice system
- mental health: terms to identify the range of mental health outcomes examined in the review.

Search terms for these three constructs were combined using the Boolean 'AND'.

Our decision to use a single comprehensive search based on these three broad constructs had the advantage that the search was not reliant on specific terms for a particular phase, which may have limited sensitivity. For example, an alternative strategy for the diagnostic test accuracy phase would be to use methodological filters to identify test accuracy studies using terms such as 'sensitivity' and 'specificity'. However, there is evidence that the inclusion of such filters in searches for such studies can lead to relevant studies being missed.[31] Another strategy would have been to list as search terms some of the more commonly used screening measures in practice and research in this area. However, this would have predetermined the type of screening measures that would be identified by the review and may have missed studies of other relevant screening measures.

The final set of search terms was developed through an iterative process. A series of pilot searches were run and the results examined and discussed by members of the research team. We considered the likely sensitivity of the search terms by establishing whether or not key citations that we knew were likely to meet inclusion criteria were retrieved by the search.

The searches were not limited by date range or language.

Databases and resources

A range of databases and resources was searched, including standard databases of predominantly peer-reviewed publications as well as resources for the identification of grey literature. The focus of the review spans the mental health literature and the criminal justice literature. We therefore specifically sought to examine databases that covered health and mental health as well as crime and social care. The following databases and resources were searched:

- PsycINFO
- MEDLINE
- EMBASE
- Cochrane Database of Systematic Reviews (CDSR)
- Database of Abstracts of Reviews of Effects (DARE)
- Cochrane Central Register of Controlled Trials (CENTRAL)
- Health Technology Assessment (HTA) database
- NHS Economic Evaluation Database (NHS EED)
- Applied Social Sciences Index and Abstracts (ASSIA)
- Criminal Justice Abstracts
- National Criminal Justice Reference Service (NCJRS)
- Social Policy & Practice
- Social Services Abstracts
- Public Affairs Information Service (PAIS) International
- Science Citation Index (SCI)
- Social Science Citation Index (SSCI)
- Conference Proceedings Citation Index – Science (CPCI-S)
- Conference Proceedings Citation Index – Social Science & Humanities (CPCI-SSH)
- Social Care Online
- The Campbell Library
- Health Economic Evaluations Database (HEED)
- OAIster
- Index to THESES
- Zetoc
- Research Papers in Economics (RePEc).

The following organisation websites and conference proceedings were also searched:

- Department of Health (www.dh.gov.uk/)
- Department for Education (www.education.gov.uk/)
- Home Office (www.homeoffice.gov.uk/)
- Joseph Rowntree Foundation (www.jrf.org.uk/)
- Royal College of Psychiatrists (www.rcpsych.ac.uk/)
- Youth Justice Board (www.yjb.gov.uk/)
- Policy Studies Institute (www.psi.org.uk/)
- Mental Health Foundation (www.mentalhealth.org.uk/)
- Young Minds (www.youngminds.org.uk/)
- Nacro (www.nacro.org.uk/)
- Revolving Doors (www.revolving-doors.org.uk/home/)
- Prison Reform Trust (www.prisonreformtrust.org.uk)
- Centre for Mental Health (www.centreformentalhealth.org.uk/index.aspx)
- British Society of Criminology (www.britsoccrim.org/)
- American Society of Criminology (www.asc41.com/).

Searches were conducted in April 2011. Full details of the specific search strategies for PsycINFO, MEDLINE and EMBASE are given in *Appendix 3*.

Additional search strategies

In addition to the searches of databases and other resources, we used three additional methods to identify relevant citations:

- Reverse citation search: we undertook reverse citation searches on all included papers using the Web of Science (WoS) Institute of Scientific Information (ISI) citation database.
- Manual check of reference lists: we conducted a manual check of the reference list of all included studies and previous major relevant reviews.
- Contact with experts: we contacted experts in the field to identify other potentially relevant papers and to request further information about included studies when necessary.

Deduplication

The number of databases searched and the use of several search strategies meant that some degree of duplication occurred. To manage this, the titles and abstracts of bibliographic records were downloaded and imported into EndNote X5 bibliographic management software (Thomson Reuters, CA, USA) and duplicate records were removed.

Screening of citations

Two reviewers screened the titles and abstracts identified in the literature search for studies that were potentially eligible for any phase of the review. Full papers of potentially eligible studies were obtained and assessed for inclusion independently by two reviewers. At both stages (first sift – titles and abstracts; second sift – full papers), disagreements were resolved by consensus or deferred to a third party if necessary.

Inclusion and exclusion criteria

We developed detailed separate PICO criteria (population/patient problem, intervention, comparison, outcome) for the different phases of the review; these are summarised in each of the relevant review chapters. Guidance was given to coders to be inclusive at the first sift (titles and abstracts) if there was any uncertainty about a citation but to apply the PICO criteria rigorously at the second sift (full papers).

Overview of the literature search

Figure 1 summarises the literature searching process. The figure given for papers identified outside of the database searches includes papers identified by website searching as well as papers identified from other sources (e.g. contact with experts). The reasons for exclusion for the studies that passed the first sift (title and abstracts) but which did not meet second sift (full paper) inclusion criteria are given in *Appendix 4*.

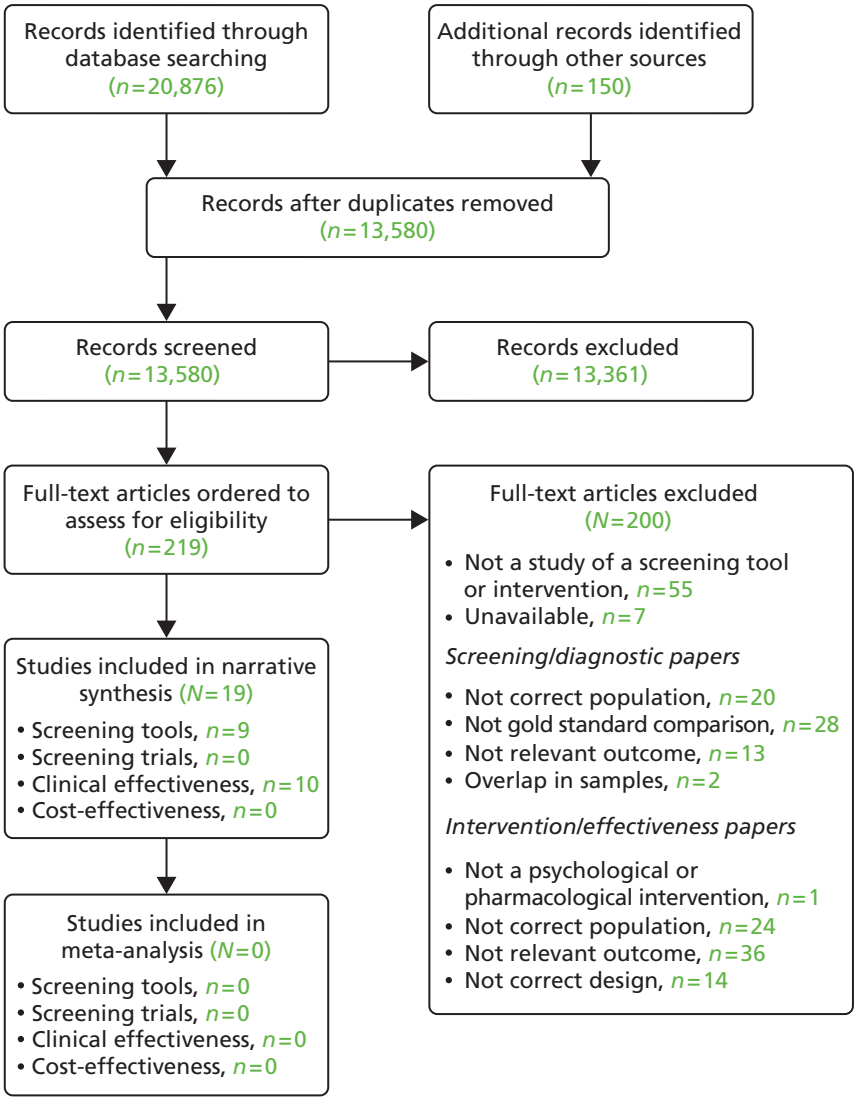


FIGURE 1 Summary of the literature search.

Chapter 4 Systematic review of diagnostic accuracy

If we are to establish whether or not screening for mental health problems in young people who offend is of benefit, a first step is to establish how accurate available screening assessments are in this population. This chapter examines the available evidence for the accuracy of different screening methods for a range of mental health problems in young people who offend. It also provides a summary of the available information on the prevalence of mental health problems according to the screening instruments identified by the review and the gold standard methods of establishing a diagnosis of a mental health problem.

Methods to assess diagnostic test accuracy

As described in *Chapter 2*, sensitivity and specificity are central concepts in understanding diagnostic test accuracy and are described in detail in *Appendix 1*, along with further information on methods of quantifying diagnostic performance.

Assessing the validity of mental health needs measures

In recognition of the argument that the presence of a diagnosis does not necessarily equate with the level of need in young people who offend,[28] we reviewed studies of screening measures designed to establish the presence of a mental health need. For these types of studies it is not possible to apply the standard strategies of assessing diagnostic accuracy because there is no gold standard of 'mental health need' against which the identification of the mental health need screening instrument can be assessed. It is impossible, therefore, to create a 2 × 2 table and summarise the performance of the screening instrument in terms of characteristics such as sensitivity and specificity.

For these studies we assessed the extent to which the assessments of mental health need had established criterion-related validity. Validity refers to the extent to which an instrument measures what it is intended to measure.[32] As applied to the question here, validity refers to whether or not a measure of mental health need in young people who offend does in fact measure the mental health needs of this group. Criterion-related validity assesses the validity of a measure by examining the extent to which it relates in ways we would expect it to relate to other measures of the same or different constructs. For example, if a mental health needs assessment is in fact a valid measure, we would expect it to relate to other indicators of mental health need, such as subsequent use of mental health services.

Rather than exclude all studies of mental health needs assessments that did not report the agreement of the measure against a gold standard diagnosis, for instruments for which we could not identify diagnostic test accuracy data we sought to include reports of validation studies that established the criterion-related validity of the mental health needs assessments.

Methods

This first phase of the review sought to answer two main questions:

1. For those screening measures reporting diagnostic status, what is their diagnostic accuracy?
2. For those screening measures identifying level of need, what evidence is there that these measures are valid indicators of mental health need?

In addition, we summarised the prevalence of mental health problems as identified by the screening instruments in these studies.

Inclusion/exclusion criteria

Two reviewers screened the titles and abstracts identified in the literature search for studies that were potentially eligible to be included in this phase of the review. Disagreements were resolved by consensus or deferred to a third party if necessary.

The PICO criteria for this stage of the review were:

- *Population and setting:* young people (aged 10–21 years) who have offended and who are in contact with the criminal justice system.
- *Intervention:* screening measures designed to identify one or more mental health diagnoses (see *Diagnostic categories*). Also included were measures that reported the presence of a mental health need. These can be brief screening measures or longer instruments. These types of measures were not diagnosis specific.
- *Reference:* for studies reporting diagnostic accuracy, a standardised diagnostic interview conducted to internationally recognised criteria [e.g. *ICD-10 Classification of Mental and Behavioural Disorders*[26] or *Diagnostic and Statistical Manual of Mental Disorders* (DSM)[27]]. For studies reporting the assessment of mental health needs, some form of validation needs to have been performed. This would typically take the form of examining the association or level of agreement between the assessment of mental health needs and one or more other indicators of mental health need.
- *Outcome:* details of the prevalence of one of the specific mental health diagnoses or mental health needs, details of the diagnostic accuracy of the measure or details of validity data for those measures reporting mental health need rather than diagnosis.
- *Study design:* cross-sectional, case–control and cohort studies and randomised controlled trials (when screening measure was used as a method of recruitment).

When citations met the inclusion criteria but reported data on samples that overlapped with those in other included studies, we examined the citations to establish whether different information on diagnostic test accuracy was reported. If so, more than one citation was included, although this was treated as a single data set. In cases in which no additional data were reported, we retained the citation reporting the largest sample size.

Diagnostic categories

For the diagnostic accuracy studies we sought evidence for a range of diagnoses, which we broadly grouped into mood disorders (e.g. major depression, bipolar disorder), anxiety disorders (e.g. generalised anxiety, panic disorder, PTSD), behavioural disruptive disorders [ADHD, conduct disorder, oppositional defiant disorder (ODD)] and a miscellaneous 'other' category that included psychotic disorder, autistic spectrum disorder and self-harm/suicide.

Although self-harm and suicidal behaviour are not diagnoses, we sought evidence of the accuracy of screening measures for these because they are important mental health outcomes with an increased prevalence in young people who offend. Unlike the diagnostic categories, for which the gold standard is typically a structured clinical interview to establish the presence of a diagnosis, for self-harm/suicide we included studies that provided details of the accuracy of the self-harm/suicide screen in terms of future self-harm or suicidal *behaviour*. Studies that assessed the screening instrument against other outcomes, such as suicidal intent, were therefore excluded.

As described earlier, we also included measures that reported the presence of a mental health need.

Although particular measures developed in the UK are recommended as screening measures, we did not presuppose that these should be prioritised in the review.

Data extraction

All data were extracted independently by two reviewers using an agreed data extraction sheet. As with the detailed PICO criteria, the data extraction sheet was first piloted on full papers and refined through an iterative process.

Quality assessment

The quality of the included studies was assessed using the QUADAS-2 tool (Quality Assessment of Diagnostic Accuracy Studies – version 2).[33] This tool examines four domains: patient selection, index test, reference standard, and flow and timing. The risk of bias is assessed for each of these domains. The first three of these also examine concerns about the applicability of the study to the review question.

The developers of the QUADAS-2 tool recommend that it is tailored to a review through the development of review-specific guidance. This may involve removing questions that are not applicable, adding additional questions that may be important quality assessment criteria for the specific subject area and providing details of how each criterion should be assessed and coded. In line with these recommendations, we developed a detailed guidance document for this review, which is given in full in *Appendix 5*.

We retained all of the risk of bias signalling questions and applicability questions. For the signalling question 'Is the reference standard likely to correctly classify the target condition?', we operationalised this as whether or not the researchers who conducted the gold standard interview had received appropriate training, had had their performance satisfactorily benchmarked or had rated well on inter-rater reliability tests. For the signalling question 'Was there an appropriate interval between the index test and the reference standard?', we defined an appropriate interval as < 2 weeks, in keeping with how this item has been applied in the evaluation of diagnostic test accuracy studies of mental health outcomes in previous versions of the QUADAS tool.[34]

The risk of bias in each domain was assessed as 'high', 'low' or 'unclear'. Concerns regarding applicability in the first three domains were also assessed as 'high', 'low' or 'unclear'.

Two reviewers independently rated the quality of the studies using the review-specific guidance. Disagreement was resolved by consensus and deferred to a third party when necessary.

Data synthesis

We produced a narrative synthesis of both the diagnostic accuracy studies and the assessment of the extent to which mental health needs screening measures are valid indicators of mental health needs in this population.

We summarised the results of the diagnostic studies in a descriptive manner. For studies that reported sufficient details to calculate 2×2 tables, we calculated sensitivity, specificity, positive likelihood ratios, negative likelihood ratios and diagnostic odds ratios (DORs) and their associated 95% confidence intervals (CIs). Analyses were conducted using Stata version 12 (StataCorp LP, College Station, TX, USA), with the diagti user-written command. For studies that reported information on diagnostic accuracy but which provided insufficient information to calculate a 2×2 table, we relied on the reports of sensitivity and specificity given in the study. There was an insufficient number of studies using the same screening measure for the same class of mental health outcomes to conduct a bivariate diagnostic meta-analysis.

Results

A total of nine studies including eight independent samples met our inclusion criteria.[35–43] Two of the included studies[38,43] reported data on samples that had some although not complete overlap with each other. The smaller of the two studies[38] reported additional details of diagnostic accuracy not reported in the larger study.[43] Specifically, Hayes *et al.*[38] reported data on the performance of a voice-administered MAYSI-2, whereas the larger study by Wasserman *et al.*[43] study reported data on a paper and pencil version alone. We therefore report the results of both studies, the larger study because of its greater size and the smaller study because of the additional information it contains on the performance of the voice-administered version of the MAYSI-2. An additional citation[44] provided a summary of the results of the included Wasserman *et al.* study.[43] All of the information contained in it was also included in the original report and so this citation was excluded. A further citation[45] reported data on a subset of a sample reported in the included Kerig *et al.* study.[39] It did not contain additional information on diagnostic test accuracy and so was also excluded in favour of the larger data set reported in Kerig *et al.*[39]

Eight of the nine studies reported data on the diagnostic test accuracy of one or more screening instrument.[35,36,38–43] The remaining study reported data on the validity of a mental health needs assessment, which will be discussed separately.[37]

Diagnostic test accuracy results

Characteristics of the included studies

A summary of the characteristics of the eight diagnostic accuracy studies is given in *Table 1*.

Setting and sample

The majority of the studies were conducted in the USA; other studies were conducted in the UK ($n = 1$[36]) and the Netherlands ($n = 1$[42]). Studies took place in a range of criminal justice settings.

Although the inclusion criteria for the review permitted samples aged between 10 and 21 years, most of the studies had a mean age of between 15 and 16 years old, with a narrow standard deviation. There was, then, a lack of representation of the diagnostic accuracy of screening instruments in the younger age group. Three of the eight studies reported data on an entirely male sample,[35,36,42] in two studies the male-to-female ratio was approximately even[38,41] and in three studies the male-to-female ratio was approximately 3 : 1.[39,40,43] Although two of the studies used overlapping samples,[38,43] the male-to-female ratio was approximately 1 : 1 in one study[38] and 3 : 1 in the other.[43] In the US studies the majority of the samples were made up of young people from a Caucasian or African American background. Ethnicity was not reported in the UK study.[36] In the Dutch study the sample was made up of those from a range of ethnic backgrounds.[42]

TABLE 1 Characteristics of the included studies

Study	Setting and sample	Screening instrument	Gold standard	Type of mental health diagnosis examined
Cashel 1998[35]	Setting: correctional facility, USA Age (years), mean (SD): 16.0 (1.0) % male: 100 Ethnicity: 46.5% African American, 26.3% white, 23.2% Hispanic American, 4.0% other *n* = 99	Instrument: MMPI-A Completion time (minutes): 90 Literacy level: seventh grade or higher (audio-administration for grades 3–6)	K-SADS-III-R (DSM-III-R)	Major depression, generalised anxiety, ADHD, conduct disorder
Grubin 2002[36]	Setting: young offender institutions, UK Age (years), range: 18–21 % male: 100 Ethnicity: not stated *n* = 30	Instrument: Prison Reception Health Screen Completion time (minutes): 5–10 Literacy level: not stated	SADS-L (RDC)	Any condition
Hayes 2005[38]	Setting: adjudicated youth, USA Age (years), mean (SD): 15.7 (1.1) % male: 52.8 Ethnicity: 56.9% African American, 39.8% white, 1.6% Hispanic, 1.6% other *n* = 123	Instrument: voice and paper MAYSI-2 Completion time (minutes): 10 Literacy level: not stated	Voice DISC (DSM-IV)	Mood disorder cluster, anxiety disorder cluster, disruptive disorder cluster
Kerig 2011[39]	Setting: county juvenile detention centres, USA Age (years), mean: 15.5 % male: 73.7 Ethnicity: 67% European American, 23% African American, 3% Hispanic, 3% multiracial, 1% American Indian/Pacific Islander and 0.5% Asian *n* = 498	Instrument: MAYSI-2 Completion time (minutes): not stated Literacy level: not stated	UCLA PTSD RI – Adolescent version (DSM-IV)	Full or partial PTSD

continued

TABLE 1 Characteristics of the included studies (*continued*)

Study	Setting and sample	Screening instrument	Gold standard	Type of mental health diagnosis examined
Kuo 2005[40]	Setting: secure facility for delinquent youth, USA Age (years), range: 13–17 % male: 74.6 Ethnicity: 51% Caucasian 34% African American *n* = 50	Instrument: MAYSI-2, MFQ, Short MFQ Completion time (minutes): 8–12 MAYSI-2; 5–7 MFQ; 2–3 Short MFQ Literacy level: not stated	Voice DISC (DSM-IV)	Depression
McReynolds 2007[41]	Setting: juvenile justice setting, USA Age (years), mean (SD): 15.7 (1.1) % male: 55.4 Ethnicity: 55.9% African American, 40.5% white, 2.1% Hispanic, 1.5% other *n* = 195	Instrument: DISC predictive scales Completion time (minutes): 15 Literacy level: third-grade oral comprehension	Voice DISC (DSM-IV)	Mood disorder cluster, anxiety disorder cluster, disruptive disorder cluster
Vreugdenhil 2006[42]	Setting: youth detention centres, the Netherlands Age (years), mean (SD): 16.4 (1.2) % male: 100 Ethnicity: 25% Dutch, 24% Surinamese, 21% Moroccan, 7% Turkish, 4% Antillean, 18% other, 2% unknown *n* = 196	Instrument: YSR Completion time (minutes): not stated Literacy level: not stated	DISC (DSM-IV)	ADHD, ODD
Wasserman 2004[43]	Setting: correctional youth setting, USA Age (years), mean (SD): 16.7 (1.5) % male: 79.7 Ethnicity: 58.2% African American, 28.3% white, 11.1% Hispanic, 2.5% other *n* = 325	Instrument: MAYSI-2 Completion time (minutes): not stated Literacy level: not stated	Voice DISC-IV (DSM-IV)	Mood disorder cluster, anxiety disorder cluster, disruptive disorder cluster

DISC, Diagnostic Interview Schedule for Children; K-SADS-III-R, Schedule for Affective Disorders and Schizophrenia for School-Age Children; MFQ, Mood and Feelings Questionnaire; MMPI-A, Minnesota Multiphasic Personality Inventory – Adolescent version; RDC, Research Diagnostic Criteria; SADS-L, Schedule for Affective Disorders and Schizophrenia – Lifetime version; SD, standard deviation; UCLA PTSD RI, University of California at Los Angeles Post-Traumatic Stress Disorder Reaction Index; YSR, Youth Self Report scale.

Screening measures used in included studies

Four studies, including three independent samples, used the MAYSI-2 as the screening instrument.[38–40,43] Kuo *et al.*[40] also examined the Mood and Feelings Questionnaire (MFQ) and a short version of the MFQ[46] in addition to the MAYSI-2. The remaining four studies each used a different screening instrument. A brief description of the screening measures used in the included studies is given below:

- *MAYSI-2.* The MAYSI-2 tool is a screening tool designed to assist juvenile justice staff in the identification of young people aged 12–17 years who may have mental health problems.[10] The tool consists of a self-report inventory of 52 questions and produces seven separate scales that focus on different areas of concern (e.g. depressed, anxious, suicidal ideation). Youths circle 'yes' or 'no' concerning whether or not each item has been true for them 'within the past few months' on six of the scales and 'ever in your whole life' on one scale. Youths can read the items themselves (the tool has a fifth-grade reading level) and circle the answers or questions can be read aloud by juvenile justice staff. A further method of administration is via a CD-ROM on a computer; youths listen to the questions using headphones and answer the questions using the keyboard or a mouse. Administration and scoring takes about 10–15 minutes.
- *Diagnostic Interview Schedule for Children (DISC) Predictive Scales (DPS).* The DPS are brief self-report measures designed to identify young people who are at increased risk of meeting diagnostic criteria for mental health difficulties.[47] The scales are derived from the DISC,[48] described in more detail in the following section, which is based on DSM criteria.[27] The scales consist of 56 items and enquire about difficulties over the last 12 months.
- *Minnesota Multiphasic Personality Inventory – Adolescent version (MMPI-A).* The MMPI-A is a self-report measure derived from the MMPI designed for adults.[49] The objective of the measure is to identify psychopathology in adolescents. The adolescent version consists of 478 items and takes approximately 90 minutes to complete. The number of items and time taken for completion mean that such a measure is unlikely to be used as a screening instrument. However, we retained the study here for two reasons. First, we did not specify a maximum completion time as part of the inclusion criteria. Second, the MMPI consists of a number of subscales, which in principle could be used as screening instruments.
- *MFQ.* The MFQ is a 33-item self-report measure based on DSM criteria[27] and designed to assess depressive symptoms in children and adolescents.[46] Items concern symptoms over the last 2 weeks and are rated as 'not true', 'sometimes true' and 'true'. The short form of the questionnaire (Short MFQ) consists of 13 items from the full scale.[46]
- *Prison Reception Health Screen.* The Prison Reception Health Screen is a 15-item measure designed to be used at intake to detect physical health, mental health and substance use disorders.[36] Slightly different versions of the scale are used for males and females, and for young people an additional item is added to identify whether or not they have experienced a recent bereavement. The instrument is designed to be administered by prison health-care staff.
- *Youth Self-Report scale (YSR).* The YSR is a standardised self-report measure for adolescents that is part of the family of measures developed by Achenbach,[50] with other measures designed for completion by parents and teachers. The scale was developed for completion by adolescents aged between 12 and 18 years. It is scored on scale from 0 ('not true') to 2 ('very true') and provides a summary of a young person's emotional and behavioural problems over the last 6 months. The scale has eight syndrome scales (e.g. anxiety and depression, somatic complaints, social problems), with a ninth scale (self-destructive/identity problems) scored for boys only and three broad problem scales (internalising, externalising, total problem score).

Gold standard instruments used in included studies

The DISC[48] was used as the gold standard in five studies, including four independent samples.[38,40–43] Two studies[35,36] used a version of the Schedule for Affective Disorders and Schizophrenia (SADS)[51] and one study[39] used the University of California at Los Angeles Post-Traumatic Stress Disorder Reaction Index (UCLA PTSD RI) – Adolescent version.[52] These three diagnostic instruments are described in more detail below:

- *DISC.* The DISC is a structured diagnostic interview to establish diagnoses for a range of mental health difficulties.[48] The interview uses a probe and follow-up format so that, if a young person answers positively to a probe question, further questions are asked to establish whether or not the person meets diagnostic criteria. The diagnoses identified by the DISC can be grouped into clusters (e.g. mood disorders, anxiety disorders, disruptive disorders). The interview takes approximately 60 minutes to complete but can be longer depending on the number of symptoms endorsed.
  The interview can be delivered in a number of formats. In the standard format the interview is administered by a trained interviewer, a delivery format used in one of the included studies.[42] An alternative format is the Voice DISC in which the young person listens to pre-recorded questions on a headphone and gives his or her response to the spoken questions using a computer keyboard. Non-clinicians, with training in the interview and computer literacy, are able to administer the Voice DISC. Four of the included studies, including three independent samples, used this format.[38,40,41,43]
  In the included studies, the accuracy of the screening instruments was typically assessed against clusters of diagnoses as determined by the DISC, including mood disorders, anxiety disorders and disruptive behavioural disorders (including ADHD).
- *SADS.* The SADS[51] is a semistructured diagnostic interview for the diagnosis of affective and psychotic disorders in adults. Responses are rated on either a 4-point scale ranging from 1 ('not at all') to 4 ('severe') or a 6-point scale ranging from 1 ('not at all') to 6 ('extreme'). It was developed before the development of DSM-III criteria and is instead based on Research Diagnostic Criteria (RDC); however, the degree of convergence between RDC and DSM diagnoses is high. The standard version asks about current mental health symptoms and the lifetime version (SADS-L) asks about previous episodes. The K-SADS-III-R (Schedule for Affective Disorders and Schizophrenia for School-Age Children) is a modified version of the SADS designed for use with children and adolescents (aged 6–18 years) and provides DSM-consistent diagnoses.[53] It uses the same 4-point and 6-point response format as the adult SADS.
- *UCLA PTSD RI – Adolescent version.* The UCLA PTSD RI – Adolescent is a 48-item measure designed to assess DSM criteria for PTSD.[52] A DSM diagnosis of PTSD requires criterion A (presence of real or perceived threat to physical integrity), criterion B (re-experiencing of traumatic event), criterion C (avoidance) and criterion D (hyper-arousal) to be met. The UCLA PTSD RI follows this structure.
  The instrument can be used to determine whether a full or partial diagnosis of PTSD is likely; a full diagnosis requires each of criterion A, B, C and D to be met; a partial diagnosis requires criterion A to be met along with two out of three of criteria B, C and D. Although the UCLA PTSD RI does not provide a formal diagnosis, we included this as a gold standard measure because it maps closely onto a recognised diagnostic system (DSM) and has convergent validity with other established gold standard diagnostic systems such as the SADS.

Quality assessment of the included studies

Table 2 summarises the risk of bias individually for the eight included studies according to QUADAS-2 criteria and *Table 3* summarises the applicability criteria individually for the eight studies. *Figures 2* and *3* provide an overall summary of the risk of bias and applicability respectively.

TABLE 2 Quality assessment of the included diagnostic test accuracy studies: risk of bias

Study	Patient selection: consecutive or random sample	Patient selection: avoided case–control	Patient selection: avoided inappropriate exclusions	Patient selection: overall risk of bias	Index test: index test interpreted blind to reference test	Index test: threshold prespecified	Index test: overall risk of bias	Reference test: reference test correctly classifies target condition	Reference test: reference test interpreted blind to index test	Reference test: overall risk of bias	Flow/timing: interval of ≤2 weeks	Flow/ timing: all participants receive a reference standard	Flow/ timing: all participants receive same reference test	Flow/ timing: all participants included in analysis	Flow/ timing: overall risk of bias
Cashel 1998[35]	?	✓	✓	Unclear	✓	?	Unclear	✓	✓	Low	?	✓	✓	✓	Unclear
Grubin 2002[36]	?	✓	✓	Unclear	✓	?	Unclear	✓	✓	Low	✓	✗	✓	✓	High
Hayes 2005[38]	✗	✓	✓	High	✓	✓	Low	✓	✓	Low	?	✓	✓	✗	High
Kerig 2011[39]	✓	✓	✓	Low	?	✗	High	✓	?	Unclear	✓	✓	✓	✓	Low
Kuo 2005[40]	✗	✓	✗	High	?	✗	High	✓	?	Unclear	?	✗	✓	✓	High
McReynolds 2007[41]	✓	✓	✓	Low	✓	✓	Low	✓	✓	Low	?	✓	✓	✗	High
Vreugdenhil 2006[42]	?	✓	✗	High	?	✓	Unclear	✓	?	Unclear	?	✗	✓	✗	High
Wasserman 2004[43]	✗	✓	✗	High	✓	✓	Low	✓	✓	Low	?	✓	✓	✗	High

✓, criterion met; ✗, criterion not met; ?, unclear if criterion met.

TABLE 3 Quality assessment of the included diagnostic test accuracy studies: applicability criteria

Study	Patient selection: applicability	Index test: applicability	Reference test: applicability
Cashel 1998[35]	Low	High	Low
Grubin 2002[36]	High	Low	Low
Hayes 2005[38]	Low	Low	Low
Kerig 2011[39]	Low	Low	Low
Kuo 2005[40]	Low	Low	Low
McReynolds 2007[41]	Low	Low	Low
Vreugdenhil 2006[42]	Low	Low	Low
Wasserman 2004[43]	Low	Low	Low

High, high level of concern about applicability; low, low level of concern about applicability.

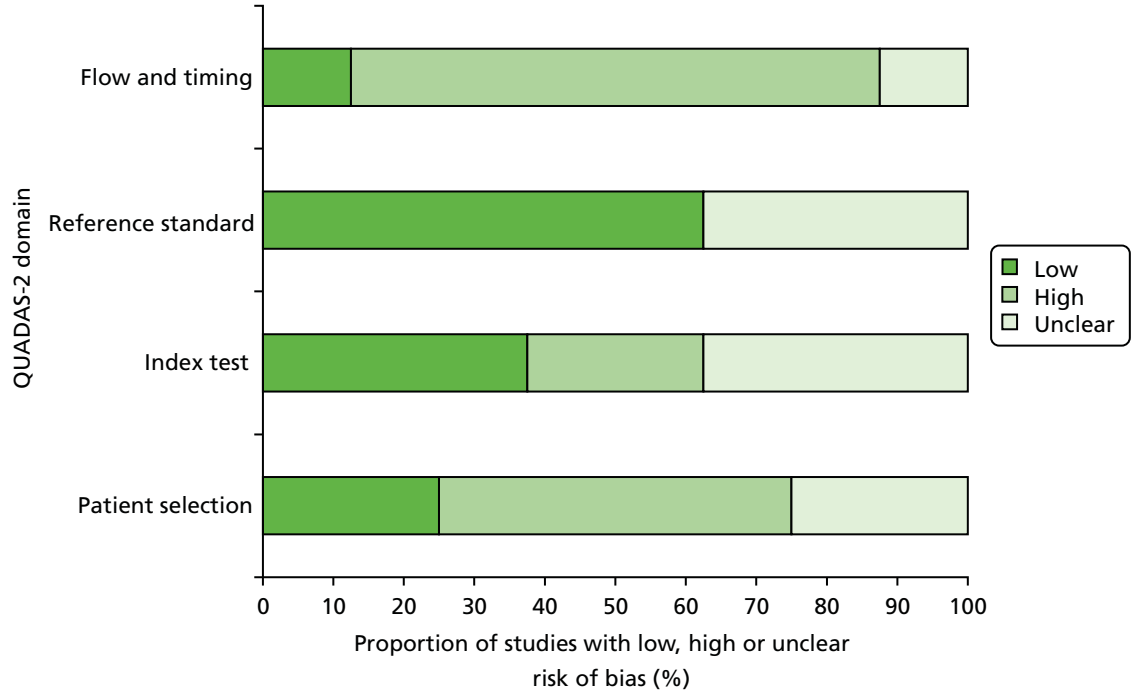


FIGURE 2 Overall risk of bias across QUADAS-2 domains for the included diagnostic test accuracy studies ($n = 8$).

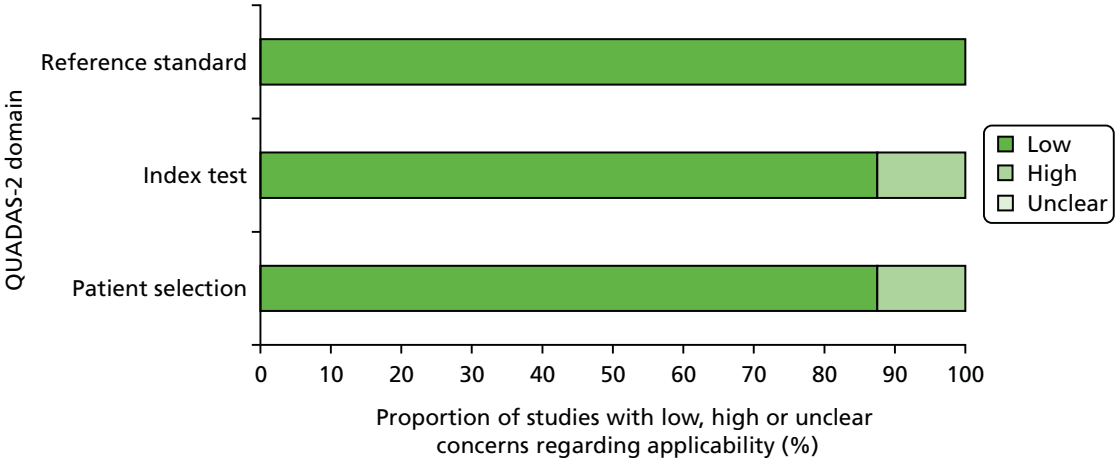


FIGURE 3 The QUADAS-2 applicability criteria for the included diagnostic test accuracy studies ($n = 8$).

Patient selection

The patient selection domain assesses if the way in which participants were selected may have introduced a bias. Four studies, consisting of three independent samples, were rated as being at high risk of bias for this domain;[38,40,42,43] the risk of bias was rated as low for two studies[39,41] and unclear for the remaining two studies.[35,36]

Although all studies avoided a case–control design, some studies did not use random or consecutive sampling for recruiting participants and others were judged to have a high number of inappropriate exclusions. The absence of random or consecutive sampling could artificially either increase or decrease the observed performance of a screening instrument against a gold standard; the direction of the influence would be determined by the exact nature of the sampling procedure used. The same is true of the high number of inappropriate exclusions from the sample. Therefore, although there is some evidence of bias for the patient selection domain, it is unclear what effect this had on the observed diagnostic accuracy in the included studies.

Index test

The index test domain asks whether the conduct or interpretation of the screening test may have introduced a bias. The overall risk of bias for this domain was rated as high for two studies,[39,40] unclear for three studies,[35,36,42] and low for three studies, consisting of two independent samples.[38,41,43]

For some studies it was unclear if the index test was interpreted blind to the reference standard. Blinding is essential to ensure that knowledge of the results does not influence the scoring of the reference standard, which may artificially inflate the observed diagnostic test accuracy of the screening test. Some studies also failed to use a prespecified cut-off point on the index test. The post hoc selection of the cut-off point can capitalise on a chance finding and artificially inflate observed diagnostic test accuracy.

Reference standard

The reference standard domain assesses whether the gold standard used or the conduct or interpretation of the gold standard test may have introduced bias. The overall risk of bias for the reference standard domain was considered low for five studies,[35,36,38,41,43] consisting of four independent samples, and unclear for three studies.[39,40,42] The unclear ratings resulted from a lack of clear evidence that the diagnostic gold standard was conducted blind to the results of the index test. As with lack of blinding for the index test, this can also distort the observed diagnostic performance of the screening test.

Flow and timing

Six out of the eight studies,[36,38,40–43] consisting of five independent samples, were rated as being at high risk of bias in terms of the flow and timing domain, which assesses whether or not the participant flow through a study and the timing of measurement may have introduced bias. The reasons for the rating of high risk for many of the studies were that not all participants received the reference standard and not all participants were included in the analysis. Participants included in the diagnostic test accuracy analysis may have differed in systematic ways from participants who were not included and this may distort the test accuracy.

Applicability criteria

Table 3 summarises by individual study the extent to which the QUADAS-2 applicability criteria are met and *Figure 3* provides an overall summary. The applicability criteria were broadly met for the patient selection and index test domains and entirely met for the reference standard domain. One study did not meet the criterion for index test applicability.[35] As we describe earlier, this was because the study used the MMPI-A, a 458-item measure taking approximately 90 minutes to complete, which makes it unsuitable as a screening measure, although one or more subscales could feasibly be used to screen. One study did not meet the applicability criterion for patient selection.[36] This was because the study recruited from a variety of adult and young offender institutions, although it was possible to extract some data for the young offender population and the results discussed later for that study are based solely on the young offender subgroup.

Summary

With the exception of the reference standard domain, for which a majority of studies had a low risk of bias, the risk of bias was either high or unclear for the majority of studies in the other three QUADAS-2 domains. No study was rated as being at low risk for all four domains. In contrast, applicability criteria were broadly met across all studies. The identified studies are therefore largely relevant to the diagnostic test accuracy question that this review seeks to answer but some caution is needed in relying on the diagnostic accuracy data reported by these studies because the level of potential bias across many QUADAS-2 domains was often unclear or high.

Results by diagnostic clusters

This section presents the diagnostic test accuracy of the included studies organised by broad diagnostic clusters. It also provides detail on the prevalence of the mental health problems as established by the screening instruments and gold standards in those samples reporting data for both types of assessment. There was an insufficient number of studies to conduct a diagnostic meta-analysis of the results for any of the broad diagnostic clusters; a narrative summary is instead given for all clusters.

When studies reported diagnostic accuracy data for multiple subscales of a measure, we report here those subscales that measure the same or a similar construct as that assessed by the gold standard. For example, Hayes *et al.*[38] report diagnostic test accuracy data for a large number of subscales of the MAYSI-2 (alcohol/drug use, angry–irritable, depressed–anxious, somatic complaints, suicidal ideation, thought disturbance), each against mood, anxiety and disruptive clusters of the DISC. Rather than report each of the subscales against the mood disorder cluster, we report the depression–anxiety subscale because of its conceptual link with the gold standard diagnosis.

Mood disorders

Five studies, consisting of four independent samples, reported information on the diagnostic test accuracy of one or more screening instrument assessed against a gold standard diagnosis of a single mood disorder or a cluster of mood disorders.[35,38,40,41,43]

Prevalence

Kuo *et al.*[40] used the MFQ and Short MFQ as a depression screen. The MFQ, using the literature standard cut-off point of 27, suggested a prevalence estimate of 24%; for the Short MFQ the figure was 42%. This compares with a prevalence figure of 14% for depression using the gold standard (Voice DISC).

McReynolds *et al.*[41] reported data on the DPS. The prevalence of any affective disorder using the DPS was 20%. A young person was classified as having an affective disorder if he or she scored positive on any of the affective predictive subscales. The prevalence for the gold standard (Voice DISC) was 11.8% for any affective disorder.

The MAYSI-2 subscales group mood and anxiety difficulties into a single subscale (depression–anxiety); therefore, it is not possible to estimate the prevalence of solely mood disorders according to standard literature cut-off points on this instrument. The figures for the depression–anxiety subscale, using the 'caution' cut-off point, are 39% for the voice-administered MAYSI-2[38] and 35.0% for the paper and pencil-administered version of the test.[43] The gold standard estimate of the prevalence of mood disorders according to the DISC affective disorder classification was 12.0% in Hayes *et al.*[38] and 10.5% in Wasserman *et al.*[43] Kuo *et al.*[40] provide data on the prevalence of depression according to what they report as the MAYSI-2 depression scale, although it is unclear if this in fact refers to the depression–anxiety subscale of the MAYSI-2.

Prevalence figures according to the MMPI used in Cashel *et al.*[35] are not given because of insufficient detail reported in that study.

Diagnostic test accuracy for mood disorders

Two studies reported diagnostic test accuracy data for major depression.[35,40] *Table 4* summarises the performance of the screening measures in these studies. Limited data are presented in *Table 4* and subsequent tables for Cashel *et al.*[35] because there was insufficient information reported in that study to calculate the 2 × 2 tables needed for the additional diagnostic test accuracy statistics.

Kuo *et al.*[40] reported literature standard cut-off points for the three screening measures examined in that study (MAYSI-2: 3; MFQ: 27; Short MFQ: 8) as well as a single alternative cut-off point for each measure. Some caution is needed in interpreting the alternative cut-off points because, unlike the literature standards that are predetermined, it is possible that the selection of these post hoc may capitalise on chance.

The sensitivity at the literature standard cut-off points for two of the instruments was in the range of 0.7–0.8, which may be unacceptably low for screening instruments because it would lead to a high proportion of people with major depression being missed. The results for the short form of the MFQ were more impressive, with a sensitivity of 1 and a specificity of approximately 0.7 at the two reported cut-off points. However, caution is needed in interpreting these results because of the small sample size and the low number of people with major depression, which means that any estimate of sensitivity is likely to be imprecise.

It is of note that, on the basis of the limited evidence presented here, the MAYSI-2, a measure designed specifically for use with young people who have offended, did not appear to have greater performance characteristics than more general measures. Although the cut-off point could be altered to increase sensitivity, this would further reduce specificity, which may lead to a high proportion of false positives. This would be problematic for the MAYSI-2, which already has low specificity at the 'caution' cut-off point; increasing sensitivity for this would further lower specificity and lead to a very high false positive rate.

Three studies, consisting of two independent samples, reported data on the diagnostic accuracy of screening instruments compared with a gold standard diagnosis of any affective disorder.[38,41,43] All three used the DISC affective disorder cluster as the gold standard. *Table 5* summarises the results for these studies. The level of sensitivity for the MAYSI-2 at the literature standard cut-off of 3 (the 'caution' cut-off) was again not as high as would be ideal for use as a screening measure. The sensitivity of the DPS was even lower at the reported cut-off point, although this was paired with a higher specificity. It is unclear if altering the cut-off point to increase the sensitivity of the DPS would retain a sufficiently high specificity to limit the number of false positives.

Summary

It is difficult to make any firm conclusions about the accuracy of screening instruments in identifying major depression or more broad affective disorder clusters as there were not enough studies estimating the accuracy of the same measures using the same cut-off points. However, on the basis of the available evidence, there is no clear indication that the performance of a measure designed specifically for use by young people who offend (MAYSI-2) is superior to that of more generic screening measures.

Anxiety disorders

Five studies, consisting of four independent samples, reported diagnostic test accuracy data on a single anxiety disorder or a cluster of anxiety disorders.[35,38,39,41,43]

Prevalence

In terms of single anxiety disorders, Kerig *et al.*[39] report a range of cut-off points on the traumatic experience subscale of the MAYSI-2 separately for males and females. There is no established cut-off point on this scale but a score of ≥ 3 has been used for research purposes. The prevalence of PTSD symptoms according to a positive screen on the MAYSI-2 using this cut-off point was 34.4% for males and 39.1% for females. According to the UCLA PTSD RI – Adolescent scale, which is treated here as the gold standard, the prevalence was in fact higher (males 49.8%; females 59.6%). There were insufficient data

TABLE 4 Diagnostic test accuracy of screening measures against a gold standard for major depression

Study	Sample size	Depressed according to gold standard (%)	Gold standard	Index test	Sensitivity (95% CI)	Specificity (95% CI)	Positive likelihood ratio (95% CI)	Negative likelihood ratio (95% CI)	DOR (95% CI)
Cashel 1998[35]	99	Not reported	K-SADS-III-R	MMPI-A scales 2, 4 and 5	0.8[a]	0.84[a]	–[a]	–[a]	–[a]
Kuo 2005[40]	50	14 ($n = 7$)	Voice DISC	MAYSI-2:					
				Cut-off: 3	0.71 (0.29 to 0.96)	0.65 (0.49 to 0.79)	2.05 (1.10 to 3.81)	0.44 (0.13 to 1.44)	4.67 (0.91 to –[b])
				Cut-off: 4	0.71 (0.29 to 0.96)	0.79 (0.64 to 0.9)	3.41 (1.62 to 7.2)	0.36 (0.11 to 1.18)	9.44 (1.76 to 49.2)
				MFQ:					
				Cut-off: 27	0.71 (0.29 to 0.96)	0.84 (0.69 to 0.93)	4.39 (1.92 to 10.0)	0.34 (0.11 to 1.11)	12.9 (2.31 to 69.6)
				Cut-off: 29	0.71 (0.29 to 0.96)	0.93 (0.81 to 0.99)	10.20 (3.12 to 33.6)	0.31 (0.09 to 0.99)	33.3 (4.95 to 225.0)
				Short MFQ:					
				Cut-off: 8	1.0 (0.59 to 1.0)	0.67 (0.52 to 0.81)	3.07 (2.0 to 4.72)	–[b]	–[b]
				Cut-off: 10	1.0 (0.59 to 1.0)	0.72 (0.56 to 0.85)	3.58 (2.22 to 5.79)	–[b]	–[b]

a Insufficient information reported in the paper to calculate the 2 × 2 table needed for the full range of diagnostic test accuracy statistics and associated CIs. Sensitivity and specificity reported here are based on the values given in the paper.
b Value/s could not be estimated.

TABLE 5 Diagnostic test accuracy of screening measures against a gold standard for any depressive disorder

Study	Sample size	Depressed according to gold standard (%)	Gold standard	Index test	Sensitivity (95% CI)	Specificity (95% CI)	Positive likelihood ratio (95% CI)	Negative likelihood ratio (95% CI)	DOR (95% CI)
Hayes 2005[38]	123	12 ($n = 16$)	DISC affective cluster	MAYSI-2 Voice: depressed/anxious (cut-off: 3)	0.75 (0.48 to 0.93)	0.71 (0.62 to 0.79)	2.59 (1.72 to 3.9)	0.35 (0.15 to 0.83)	7.35 (2.30 to 23.3)
McReynolds 2007[41]	195	13.8 ($n = 27$)	DISC affective cluster	DPS: any affective	0.57 (0.35 to 0.77)	0.85 (0.79 to 0.9)	3.74 (2.26 to 6.19)	0.51 (0.32 to 0.82)	7.3 (2.95 to 18.1)
Wasserman 2004[43]	325	10.5 ($n = 34$)	DISC affective cluster	MAYSI-2 paper and pencil: depressed/anxious (cut-off: 3)	0.74 (0.56 to 0.87)	0.60 (0.54 to 0.66)	1.84 (1.44 to 2.36)	0.44 (0.25 to 0.78)	4.19 (1.92 to 9.14)

reported in Cashel *et al.*[35] to calculate prevalence estimates for generalised anxiety disorder according to the screening instrument or the gold standard instrument.

In the study by McReynolds *et al.*,[41] the prevalence of anxiety disorders according to a positive score on any of the anxiety DPS was 65.6%. For the gold standard measure (DISC), the prevalence of any anxiety disorder ranged from 21.2% to 27.6% (see *Table 7*).

Diagnostic test accuracy for anxiety disorders

Two studies reported data on the diagnostic performance of screening measures for a single anxiety disorder as established by a gold standard. Cashel *et al.*[35] report data for generalised anxiety and Kerig *et al.*[39] report data for full or partial PTSD. *Table 6* summarises the results of these two studies. Cashel *et al.*[35] examined a number of MMPI-A scales as a screen for generalised anxiety disorder and reported a sensitivity of 0.73 and a specificity of 0.84. There were insufficient data reported in this study to extract a 2 × 2 table and so CIs and other diagnostic performance characteristics could not be calculated. As described earlier, there is no agreed cut-off point for the MAYSI-2 traumatic experience subscales, as used in Kerig *et al.*[39] At the cut-off point of 3, which has been used for research purposes, the MAYSI-2 traumatic experience subscale had modest sensitivity and good specificity for both the males and females in the sample (see *Table 6*).

The same three studies that reported data on the diagnostic accuracy of screening measures for any depressive disorder (including two independent samples) also assessed their accuracy against a gold standard measure of 'any anxiety disorder'.[38,41,43] All three used the DISC anxiety disorder cluster as the gold standard. *Table 7* summarises the results for these studies. The Hayes *et al.*[38] and Wasserman *et al.*[43] studies, which had overlapping samples, reported modest sensitivity for the MAYSI-2 at the literature standard cut-off points, combined with modest specificity. The McReynolds *et al.*[41] study used the DPS as the screening measure. Participants were scored positively if they scored above the cut-off on any of the anxiety scales. Sensitivity was very high (0.97, 95% CI 0.88 to 0.99) but this was combined with low specificity (0.44, 95% CI 0.36 to 0.52).

Summary

There were too few studies to conduct a diagnostic meta-analysis or make firm conclusions about the diagnostic performance of any of the instruments in the identification of anxiety disorders among young people who offend. As with the results for depressive disorders, there is not a clear indication that the MAYSI-2, a test specifically designed for young people who have offended, has superior operating characteristics relative to other more general screening instruments.

Disruptive disorders

Five studies,[35,38,41–43] consisting of four independent samples, reported data on the diagnostic accuracy of screening instruments for disruptive disorders.

Prevalence

In terms of specific disruptive disorders, prevalence estimates could not be calculated for the Cashel *et al.*[35] study because insufficient information was reported to carry out the calculations. The prevalence of ADHD according to the acceptable sensitivity cut-off point of the attention deficit hyperactivity (ADH) subscale of the YSR was 53.1% in Vreugdenhil *et al.*;[42] the prevalence according to the gold standard (DISC) was 8%. Vreugdenhil *et al.*[42] also report data on ODD. If the aggressive subscale of the YSR is used to estimate the prevalence of ODD it suggests a figure of 85.7%; when the externalising subscale is used the figure is 53.1%. The gold standard suggests a figure of 14%.

The prevalence of any disruptive disorder according to the angry–irritable subscale (cut-off 5) of the MAYSI-2 (voice version) was 44.7% according to Hayes *et al.*[38] For the paper and pencil version, the figure was 38.5%.[43] McReynolds *et al.*[41] used the DPS as the screening instrument; a positive score on any of the disruptive scales gave a prevalence estimate of any disruptive disorder of 51.3%. Gold standard estimates

TABLE 6 Diagnostic test accuracy of screening measures against a gold standard for a single anxiety disorder

Study	Sample size	Anxiety disorder according to gold standard (%)	Gold standard diagnosis	Index test	Sensitivity (95% CI)	Specificity (95% CI)	Positive likelihood ratio (95% CI)	Negative likelihood ratio (95% CI)	DOR (95% CI)
Cashel 1998[35]	99	Generalised anxiety: % not reported	K-SADS-III-R	MMPI-A scales 2, 4 and 6	0.73[a]	0.84[a]	–[a]	–[a]	–[a]
Kerig 2011,[39] male sample	337	Full or partial PTSD: 49.9 ($n = 168$)	UCLA PTSD RI – Adolescent scale	MAYSI-2 traumatic subscale (cut-off: 3)	0.54 (0.46 to 0.62)	0.85 (0.79 to 0.90)	3.66 (2.48 to 5.40)	0.54 (0.45 to 0.64)	6.81 (4.05 to 11.4)
Kerig 2011,[39] female sample	161	Full or partial PTSD: 59.6 ($n = 96$)	UCLA PTSD RI – Adolescent scale	MAYSI-2 traumatic subscale (cut-off: 3)	0.55 (0.44 to 0.65)	0.85 (0.74 to 0.92)	3.59 (1.97 to 6.53)	0.53 (0.41 to 0.68)	6.78 (3.12 to 14.7)

a Insufficient information reported in the paper to calculate the 2 × 2 table needed for the full range of diagnostic test accuracy statistics and associated CIs. Sensitivity and specificity reported here are based on the values given in the paper.

TABLE 7 Diagnostic test accuracy of screening measures against a gold standard for any anxiety disorder

Study	Sample size	Any anxiety disorder according to gold standard (%)	Gold standard	Index test	Sensitivity (95% CI)	Specificity (95% CI)	Positive likelihood ratio (95% CI)	Negative likelihood ratio (95% CI)	DOR (95% CI)
Hayes 2005[38]	123	27.6 ($n = 34$)	DISC anxiety cluster	MAYSI-2 Voice: depressed/ anxious (cut-off: 3)	0.52 (0.35 to 0.7)	0.73 (0.62 to 0.81)	1.93 (1.22 to 3.05)	0.65 (0.44 to 0.95)	2.97 (1.32 to 6.66)
McReynolds 2007[41]	195	22.6 ($n = 44$)	DISC anxiety cluster	DPS: any anxiety	0.97 (0.88 to 0.99)	0.44 (0.36 to 0.52)	1.74 (1.50 to 2.01)	0.05 (0.01 to 0.36)	33.4 (5.65 to –[a])
Wasserman 2004[43]	325	21.2 ($n = 69$)	DISC anxiety cluster	MAYSI-2 paper and pencil: depressed/ anxious (cut-off: 3)	0.70 (0.57 to 0.8)	0.64 (0.58 to 0.7)	1.94 (1.54 to 2.43)	0.48 (0.33 to 0.69)	4.07 (2.31 to 7.19)

a Value could not be estimated.

for the three studies, all of which used the DISC as the gold standard, ranged from 28.6%[43] to 39%,[38] although these two studies had overlapping samples.

Diagnostic test accuracy for disruptive disorders

Two studies report separate accuracy estimates for specific disruptive disorders: Cashel *et al.*[35] provide data for ADHD and conduct disorder; Vreugdenhil *et al.*[42] report data for ADHD and ODD (*Table 8*). For the study by Cashel *et al.*,[35] as with the description of the results for anxiety and depressive disorders, the reported diagnostic information for disruptive disorders is limited to that given in the paper, because there was insufficient information to extract 2 × 2 tables. Cashel *et al.*[35] used various MMPI scales to screen for ADHD and conduct disorder against the gold standard K-SADS-III-R. Although the results for conduct disorder suggested a combination of modest sensitivity and specificity, the results for ADHD were somewhat higher (sensitivity 0.77; specificity 0.84) (see *Table 8*).

The same three studies that reported data on the diagnostic accuracy of screening measures for any depressive disorder and any anxiety disorder (consisting of two independent samples) also assessed their accuracy against a DISC diagnosis of any disruptive disorder.[38,41,43] *Table 9* summarises the results for these studies. As with the results for anxiety and depression, the MAYSI-2 reported modest sensitivity and specificity for the prediction of any disruptive disorder. The McReynolds *et al.*[41] study reported good sensitivity (0.89, 95% CI 0.79 to 0.96) and modest specificity (0.67, 95% CI 0.58 to 0.75) for the DPS. In this study, participants were scored positively if they scored above the cut-off on any of the disruptive scales.

Summary

The results for disruptive disorders are similar to those for anxiety and depressive disorders. There were too few studies to make firm conclusions about the diagnostic test accuracy of any screening measure, and the results for the MAYSI-2 indicated a combination of modest sensitivity and modest specificity.

Other mental health problems

In addition to mood disorders, anxiety disorders and disruptive disorders, we searched for evidence on the diagnostic accuracy of screening measures for a number of additional mental health problems, including psychosis and autistic spectrum disorders. We also searched for studies examining the capacity of screening measures to identify subsequent self-harm and suicidal behaviour. We found no studies that met inclusion criteria for these mental health problems.

The study by Grubin *et al.*[36] met inclusion criteria but has not been discussed so far because the outcome of interest was 'any mental health condition', rather than a specific disorder or cluster of disorders as used in this chapter to group studies.

Grubin *et al.*[36] examined the effectiveness of 'new prison reception health screening arrangements' in identifying physical and mental health needs at 10 prisons in the UK, of which two were young offenders institutions housing young men aged 18–21 years. Young women aged 16–21 years were also included in the study; however, it was not possible to extract data separately for the young women and so data reported here are for the sample of young men only.

The health screening instrument contained 15 basic screening questions and was administered on reception to the facility. At each prison diagnostic interviews were carried out with a random sample of 15 prisoners using the SADS-L. Validation data were therefore available for 30 young male offenders from the two young offender institutions. Although the study report contains sufficient data to extract a 2 × 2 table, data are not reported here on sensitivity and specificity because only two young people met criteria for a mental health problem, which makes it difficult to provide a meaningful estimate of sensitivity.

TABLE 8 Diagnostic test accuracy of screening measures against a gold standard for specific disruptive disorders

Study	Sample size	Disruptive disorder according to gold standard (%)	Gold standard	Index test	Sensitivity (95% CI)	Specificity (95% CI)	Positive likelihood ratio (95% CI)	Negative likelihood ratio (95% CI)	DOR (95% CI)
Cashel 1998[35]	99	% not reported	K-SADS-III-R:	MMPI-A:					
			(1) ADHD	Scales 7 and 9	0.77[a]	0.84[a]	–[a]	–[a]	–[a]
			(2) Conduct disorder	Scales 2–4 and 6	0.60[a]	0.55[a]	–[a]	–[a]	–[a]
Vreugdenhil 2006[42]	196	ADHD: 8 ($n = 16$)	DISC: (1) ADHD	YSR:					
				ADH maximum sensitivity	1.0 (0.79 to 1.0)	0.33 (0.26 to 0.40)	1.49 (1.34 to 1.65)	–[b]	–[b]
				ADH acceptable sensitivity	0.88 (0.62 to 0.98)	0.50 (0.43 to 0.58)	1.75 (1.38 to 2.22)	0.25 (0.07 to 0.92)	7.0 (1.72 to –[b])
		ODD: 14 ($n = 27$)	DISC: (2) ODD	YSR:					
				Aggressive behaviour	0.96 (0.81 to 0.99)	0.16 (0.11 to 0.22)	1.15 (1.04 to 1.27)	0.23 (0.03 to 1.64)	4.94 (0.81 to –[b])
				Externalising problems	0.93 (0.76 to 0.99)	0.20 (0.14 to 0.26)	1.15 (1.01 to 1.31)	0.38 (0.10 to 1.49)	3.03 (0.76 to –[b])

a Insufficient information reported in the paper to calculate the 2 × 2 table needed for the full range of diagnostic test accuracy statistics and associated CIs.
b Value/s could not be estimated.

TABLE 9 Diagnostic test accuracy of screening measures against a gold standard for any disruptive disorder

Study	Sample size	Any disruptive disorder according to gold standard (%)	Gold standard	Index test	Sensitivity (95% CI)	Specificity (95% CI)	Positive likelihood ratio (95% CI)	Negative likelihood ratio (95% CI)	DOR (95% CI)
Hayes 2005[38]	123	39 ($n = 43$)	DISC disruptive disorder cluster	MAYSI-2 Voice: angry–irritable (cut-off: 5)	0.65 (0.49 to 0.79)	0.66 (0.55 to 0.76)	1.93 (1.32 to 2.81)	0.52 (0.34 to 0.82)	3.66 (1.69 to 7.94)
McReynolds 2007[41]	195	32.8 ($n = 64$)	DISC disruptive disorder cluster	DPS: any disruptive	0.89 (0.79 to 0.96)	0.67 (0.58 to 0.75)	2.71 (2.09 to 3.52)	0.16 (0.08 to 0.33)	16.7 (7.13 to 38.8)
Wasserman 2004[43]	325	28.6 ($n = 93$)	DISC disruptive disorder cluster	MAYSI-2 paper and pencil: angry–irritable (cut-off: 5)	0.57 (0.46 to 0.67)	0.69 (0.49 to 0.8)	1.83 (1.41 to 2.37)	0.63 (0.49 to 0.8)	2.93 (1.79 to 4.8)

Validity of the mental health needs assessment results

Some screening measures for mental health problems in young people who offend do not provide sensitivity and specificity against a gold standard diagnosis; instead, they aim to identify a 'mental health need'. For any screening measure described as a mental health needs assessment and for which there were no diagnostic test accuracy data available, we sought validation studies of that screen as a measure of mental health need. As described in more detail earlier, for inclusion studies had to provide evidence of some form of criterion-related validity for a mental health need.

Although there are a number of screening measures that are designed to identify mental health needs in young people who offend, including a number developed in the UK, we identified only one study that met inclusion criteria.[37]

Haapanen and Steiner[37] assessed the performance of a battery of tools termed the Mental Health and Substance Abuse Treatment Needs Assessment. Although this battery of tests includes some measures for which there exist diagnostic test accuracy data, such as the MAYSI-2, we included this study because it was the entire battery of tests that was designed to assess mental health needs. The full battery consisted of the Achenbach Child Behavior Checklist – YSR, the MAYSI-2, the Weinberger Adjustment Inventory and the Drug Experience Questionnaire. Paper and pencil versions of the test were used.

The sample consisted of 836 young people who were all committed to the California Youth Authority, which deals with young people who have committed very serious crimes, who have a substantial criminal history or who have failed at local interventions. In total, 79.4% of the sample was male. The average age of the sample was described as 16–17 years and the ethnicity of the sample was described as predominantly Hispanic or African American. Further details about ethnicity were not given. Validation of this combined mental health assessment was based on a case-note review, which was used to establish whether the young people were offered mental health treatment, were offered psychopharmacological treatments or were identified as requiring treatment but treatment was not provided. The level of reporting of the results is limited and the results are largely descriptive. In general, however, the authors state that elevated scores on instruments such as the MAYSI-2 and YSR were related to an increased use of mental health services or need for such services, at least for the male sample.

Although the search identified a number of studies reporting data on measures used as part of the Asset screening pathway in the UK [e.g. Asset, Screening Questionnaire Interview for Adolescents (SQiFA), SIfA], none of the studies met our inclusion criteria for establishing the criterion validity of the screening measure against other measures of mental health need. For example, validity data are reported for the Asset instrument[54] but, as this instrument is designed primarily to identify the factors contributing to a young person's offending, the validation was against indices of reoffending rather than mental health need.

Another report described a number of studies, one of which examined the use of the Salford Needs Assessment Scale for Adolescents (SNASA) in 301 young people who had offended and also interviewed the case managers of the sample to enquire about needs.[4] However, after discussion we concluded that this comparator – the views of case managers – did not constitute adequate evidence of criterion-related validity. A further citation reported data on the reliability of the SNASA but did not report data directly relevant to assessing criterion-related validity.[55]

It should be recognised, however, that our search strategy may have missed important studies in this respect. For a study to be identified, it had to mention the measure in the title or abstract along with information about validity. It is possible that studies may contain information relevant to establishing the validity of a measure without the measure being referred to in the title and abstract. For example, studies that used one of the mental health needs assessments as the outcome measure in a trial would provide evidence relevant to establishing criterion-related validity, but such studies would not necessarily be identified as part of the search if the measure was not referred to in the abstract. For example, the SNASA

was used as an outcome measure in a trial identified as part of the clinical effectiveness review.[56] Full details of this study are given in *Chapter 6*. In brief, the SNASA was used in a trial of cognitive–behavioural therapy (CBT) compared with treatment as usual, with the intervention designed to improve a range of mental health outcomes, including depression and anxiety symptoms. Outcome was assessed at 11 months' follow-up. The two groups appeared to be broadly comparable at follow-up on the SNASA [CBT group ($n = 18$): mean 10.5, standard deviation (SD) 3.54)]; treatment as usual group ($n = 20$): mean 10.75, SD 4.0]. Evidence for criterion-related validity would have required lower scores at follow-up in the CBT group. However, the absence of such a relationship and its implications for understanding the validity of the SNASA should be treated with caution given the small sample size.

Summary

There were too few studies to make any firm conclusions about the diagnostic test accuracy of any of the screening measures examined in this review. Ideally, a conclusion about a particular screening measure would require a large number of studies reporting diagnostic performance at a range of cut-off points. This would allow the calculation of pooled estimates of sensitivity, specificity and other indices of test accuracy. It would also allow the examination of potential sources of heterogeneity in observed test accuracy across studies.

It is also difficult to draw conclusions about the comparative performance of different screening methods, even though some studies examined the performance of more than one measure in the same sample. Studies did not typically compare the performance of the measures across the full range of potential cut-off points. In some cases the reported performance of one measure had a particular balance between sensitivity and specificity whereas another measure had a different balance. It is unclear how the two measures would compare at cut-off points that attempted to give broadly the same balance (e.g. reasonably high sensitivity combined with acceptable specificity).

As discussed in *Chapter 2*, in which the decision problem was outlined, one of the aims of screening may be to detect previously unidentified cases. The diagnostic test accuracy data reported by the studies included in this review did not appear to differentiate between previously identified and unidentified cases. It is unclear, therefore, how the reported accuracy would be altered if the analysis was restricted to previously unidentified cases.

Any conclusions about the diagnostic performance of screening measures in populations of young people who have offended are, therefore, necessarily tentative. One potential conclusion is that there is no clear evidence that the MAYSI-2, a test specifically designed for use in this population, has superior operating characteristics to those of other general measures. More generally, the reported literature standard cut-off points for the MAYSI-2 and other instruments suggest a combination of both moderate sensitivity and moderate specificity, and so altering the cut-off point to increase sensitivity may lead to unacceptably low specificity.

One of the objectives of the review was to establish in which groups of young offenders screening may be of use. There were too few studies to make any firm conclusions about such groups, for example community compared with incarcerated settings or particular diagnostic subgroups. However, data were identified on screening accuracy for some mental health problems, including depression, PTSD, ADHD, conduct disorder and ODD. These disorders are therefore candidates for use as exemplars in the decision model.

Reflections on policy and practice

Although current UK policy recommends screening for mental health problems in young people who offend, there is currently a limited evidence base examining the diagnostic test accuracy of available screening measures.

There also appears to be limited validation of mental health needs assessments. Although this information would be useful, it is not immediately clear how it could be used to inform decision-making in terms of the clinical effectiveness and cost-effectiveness of a screening strategy, because this typically requires sufficient data to calculate sensitivity and specificity against a gold standard.

The majority of the diagnostic test accuracy studies were conducted in the USA; only one was conducted in the UK. Furthermore, many of the studies were conducted in incarcerated settings. In the UK, in contrast, most young people who offend are managed in the community. The existing literature – already limited in terms of the number and quality of studies – may therefore be further limited by the extent to which the findings can be generalised to UK settings.

Chapter 5 Clinical effectiveness of screening strategies

Chapter 4 examined the accuracy of the available screening measures for mental health problems in young people who have offended, a first step in establishing whether or not screening is a valuable strategy. A next step is to examine the clinical effectiveness of these screening strategies. In other words, when a screening strategy is implemented to identify young people with mental health problems, whether alone or in combination with enhanced care or treatment, what is the impact of this on mental health outcomes?

In this next stage of the review we sought to identify any randomised controlled trials or non-randomised controlled trials that had examined the effect on mental health outcomes of screening strategies in young people who have offended.

Overview of screening study designs

There are a number of potentially relevant screening designs. In the simplest design, participants are allocated to either a screening condition or a no-screening condition, in which a formal screening method is used to establish whether or not someone meets predetermined criteria for having a particular mental health problem. The impact of the screening intervention on relevant mental health outcomes is then assessed. Other designs link the results of the screening to some form of enhancement of care. For example, participants could be assigned to a screening or a no-screening condition and those people in the screening condition who score above a clinical cut-off point could then be offered some form of enhanced care (e.g. further assessment followed, if necessary, by a pharmacological or psychological treatment). All designs were considered relevant and we sought to identify any trials that had used these approaches in a young offender population.

Method

In this phase of the review we sought to answer the following question: 'What is the clinical effectiveness of screening strategies for mental health problems in young people who offend?'

The search strategy outlined in *Chapter 3* was used to identify relevant studies.

Inclusion and exclusion criteria

The PICO criteria used to guide the selection were:

- *population and setting:* young people (aged 10–21 years) who have offended and who are in contact with the criminal justice system
- *intervention:* implementation of a screening strategy
- *comparison:* usual care
- *outcome:* mental health outcomes over the short or long term
- *study design:* randomised controlled trials or non-randomised controlled trials.

We chose to include non-randomised controlled trials as well as randomised controlled trials because our initial scoping of the literature identified a lack of evidence in this area. In an effort to identify all relevant work we therefore included additional designs even though these may have greater threats to internal validity.

Data extraction

All studies were independently examined by two reviewers and any discrepancies were resolved by consensus or deferred to a third party if necessary.

Results

Despite an extensive search involving a large number of databases and additional searches of other resources, resulting in the examination of > 13,000 citations, we identified no studies that met the PICO criteria.

Summary

In the absence of evidence of the clinical effectiveness of screening strategies for mental health outcomes we undertook a further review, described in the next chapter, in which we sought to identify any evidence of the clinical effectiveness of interventions for mental health difficulties.

Reflections on policy and practice

Screening for mental health problems is currently recommended for young people who have offended. However, to date there are no trials examining the effectiveness of such a strategy.

Chapter 6 Clinical effectiveness of treatments for mental health difficulties

Our initial scoping of the literature had anticipated that we might identify few if any studies that had examined the clinical effectiveness of screening strategies for mental health difficulties in young people who have offended. We therefore also sought to identify the wider literature on the effectiveness of treatments for mental health difficulties in this group of young people. This chapter provides the results of this further systematic review of clinical effectiveness.

Method

In this phase of the review we sought to answer the following research question: 'What is the clinical effectiveness of interventions for mental health problems for young people who have offended?'

Inclusion and exclusion criteria

Two reviewers screened the titles and abstracts identified in the literature search for studies that were potentially eligible to be included in this phase. Any disagreements were deferred to a third party. Full papers of potentially eligible studies were obtained and assessed for inclusion independently by two reviewers. Disagreements were resolved by consensus or deferred to a third party if necessary.

The PICO criteria used to guide study selection were:

- *population and setting:* young people (aged 10–21 years) who have offended and who are in contact with the criminal justice system
- *intervention:* any psychological or pharmacological intervention for the mental health problems included in this review
- *comparison:* other active treatment, no intervention, placebo, attention control or other 'psychological placebo', or usual care
- *outcome:* primary outcomes relating to the mental health problem targeted by the intervention; secondary outcomes include other mental health problems, quality of life measures, educational attainment and further contact with the care and criminal justice system
- *study design:* randomised controlled trials.

In terms of the mental health problems targeted by the interventions, we sought evidence for any of the categories described in *Chapter 4*. This included mood disorders (e.g. major depression, bipolar disorder), anxiety disorders (any anxiety disorder), disruptive disorders (including ADHD) and a small number of other mental health difficulties including autistic spectrum disorder, psychotic presentations and self-harm or suicidal behaviour.

This review focused exclusively on studies that reported outcomes relevant to these mental health problems. If a study solely reported outcomes that may be linked in some way to mental health (e.g. self-esteem) but did not report at least one outcome measuring one of the specified mental health constructs, it was not included. Studies that reported the effectiveness of an intervention for recidivism but not mental health problems were excluded.

Data extraction

Data were extracted independently by two researchers to a standardised data extraction form. This included characteristics of the intervention (e.g. type, duration, health professional involvement) and primary and secondary outcomes (e.g. change on continuous mental health measures, dichotomous change in diagnostic status).

Quality assessment

We judged methodological quality and sources of bias using the Cochrane risk of bias tool.[57] This tool contains seven domains (e.g. random sequence generation) that correspond to areas of potential bias in a randomised controlled trial. The risk of bias in each domain was assessed for each study and given a rating of 'high', 'low' or 'unclear'. Together, these seven domains provided a picture of the overall risk of bias for each study.

Data synthesis

Our foreknowledge of the literature suggested that there may be an insufficient number of studies using comparable interventions for comparable psychological difficulties to conduct a meta-analysis. This was the case. We therefore produced a narrative synthesis of the results and facilitated cross-comparisons of effectiveness by calculating a measure of effect and associated 95% CI (continuous measures: Cohen's *d*; dichotomous measures: relative risk). For the standardised mean difference (Cohen's *d*; SMD), scores of < 0 favour treatment; for relative risk, scores of < 1 favour treatment. Analyses were conducted in Stata version 12 using the metan function. When studies did not report the information typically required to calculate effect sizes (e.g. mean, SD, sample size for SMDs), the strategies for estimating effect sizes in the absence of complete information outlined in the *Cochrane Handbook for Systematic Reviews of Interventions* were used when possible.[58]

Results

Characteristics of the included studies

We identified 10 studies that met the inclusion criteria;[56,59–67] the characteristics of the included studies are summarised in *Table 10*. All of the trials were of a psychological intervention; no studies of pharmacological treatments met our inclusion criteria.

TABLE 10 Characteristics of the included studies

Study	Setting and sample	Intervention/s	Comparator	Main treatment focus and outcomes
Ahrens 2002[59]	Setting: facility for adolescent offenders, USA Age (years), mean (SD): 16.4 (not reported) % male: 100 Ethnicity: 60.5% Caucasian, 26.3% African American, 5.3% Hispanic, 5.3% Native American, 2.6% other	Cognitive processing therapy (eight 60-minute group sessions delivered by a doctoral candidate and psychologist) ($n = 19$)	Wait-list control ($n = 19$)	Treatment focus: traumatic symptoms Outcomes measured: trauma (PSS-SR, IES), depression (BDI) Measurement point: post treatment
Biggam 2002[60]	Setting: young offender institution, UK Age (years), mean (SD): 19.3 (1.3) % male: not reported Ethnicity: not reported	Problem-solving therapy (five 90-minute group sessions delivered by a psychologist) ($n = 23$)	No-treatment control ($n = 23$)	Treatment focus: symptoms related to stress of incarceration, including depression and anxiety Outcomes measured: depression (HADS), anxiety (HADS) Measurement points: post treatment, 3 months post treatment

TABLE 10 Characteristics of the included studies (*continued*)

Study	Setting and sample	Intervention/s	Comparator	Main treatment focus and outcomes
Martsch 2005[61]	Setting: court ordered to attend treatment in a community mental health agency, USA Age (years), mean (SD): 15.9 (not reported) % male: 100 Ethnicity: 83.1% Caucasian, 13.8% African American, 3.1% Hispanic	(1) CBT (10 × 120-minute group sessions delivered by two MSW-level therapists) ($n = 31$; $n = 14$ at follow-up) (2) CBT + focus on group process (10 × 120-minute group sessions delivered by two MSW-level therapists) ($n = 34$; $n = 17$ at follow-up)	Not applicable	Treatment focus: aggressive behaviour, including conduct disorder Outcome measured: conduct disorder (subscale of the RBPC) Measurement points: post treatment, 9 months post treatment
Mitchell 2011[56]	Setting: four children's secure homes and one young offender institution, UK Age (years), mean (SD): 15.6 (1.6) % male: 100 Ethnicity: 97.5% white	CBT (10 one-to-one sessions, duration not stated, delivered by a mental health practitioner) ($n = 19$; $n = 18$ at follow-up)	TAU ($n = 21$; $n = 20$ at follow-up)	Treatment focus: range of mental health problems Outcomes measured: depression (DCP), anxiety (DCP) Measurement point: 11 months' follow-up
Persons 1966[62]	Setting: state reformatory, USA Age (years), mean (SD): 16.4 (not reported) % male: 100 Ethnicity: 80.5% white, 19.5% African American	Psychotherapy (40 × 90-minute group sessions and 20 × 60-minute one-to-one sessions delivered by psychotherapists) ($n = 41$)	TAU ($n = 41$)	Treatment focus: interpersonal relationships Outcomes measured: anxiety (MAS), depression (MMPI), psychasthenia (MMPI) Measurement point: post treatment
Reardon 1976[63]	Setting: treatment facility for adolescent delinquent females, USA Age (years), mean (SD): 16 (not reported) % male: 0 Ethnicity: not reported	(1) CBT: rational stage-directed imagery (six 60-minute one-to-one sessions delivered by doctoral students) ($n = 8$) (2) CBT: rational stage-directed therapy (six 60-minute one-to-one sessions delivered by doctoral students) ($n = 8$)	(1) Attention control (six 60-minute one-to-one sessions) ($n = 8$) (2) No-treatment control ($n = 8$)	Treatment focus: reduce psychological stress Outcomes measured: psychosis (TSCS), neurosis (TSCS), anxiety (MAACL), depression (MAACL) Measurement points: post treatment, 2 months post treatment
Rohde 2004[64]	Setting: County Juvenile Justice Department, USA Age (years), mean (SD): 15.1 (1.4) % male: 51.63 Ethnicity: 80.6% white	CBT: Adolescent Coping with Depression course (16 × 120-minute group sessions delivered by mental health practitioners) ($n = 45$)	'Life skills' (attention control) (16 × 120-minute group sessions) ($n = 48$)	Treatment focus: depression comorbid with conduct disorder Outcomes measured: depression (LIFE, K-SADS, HRSD, BDI-II), externalising (YSR) Measurement points: post treatment, 6 months and 12 months post treatment

continued

TABLE 10 Characteristics of the included studies (*continued*)

Study	Setting and sample	Intervention/s	Comparator	Main treatment focus and outcomes
Rohde 2004[65]	Setting: youth correctional facility, USA Age (years), mean (SD): 16.3 (1.9) % male: 100 Ethnicity: 64.2% white, 14.2% Latino, 10.4% Native American, 6.7% African American, 2.2% Asian, 0.7% Pacific Islander, 1.5% other	CBT coping course (16 × 90-minute group sessions delivered by mental health practitioners) ($n = 46$)	TAU ($n = 30$)	Treatment focus: general coping Outcomes measured: internalising (YSR), externalising (YSR) Measurement point: post treatment
Scherer 1994[66]	Setting: community setting, USA Age (years), mean (SD): 15.1 (not reported) % male: 81.8 Ethnicity: 78% African American, 22% white	Multisystemic Family Preservation Program (multiple weekly meetings delivered by mental health professionals) ($n = 23$)	TAU ($n = 21$)	Treatment focus: multiple determinants of juvenile delinquency Outcomes measured: conduct disorder (RBPC), depression (BSI) Measurement point: post treatment
Shivrattan 1988[67]	Setting: incarcerated adolescents, Canada Age (years), range: 15–17 % male: 100 Ethnicity: not reported	(1) CBT: Social Interaction Skills Program (eight 60-minute one-to-one treatment sessions delivered by a psychology graduate or teacher) ($n = 15$; $n = 14$ at follow-up) (2) Behaviour therapy: Stress Management Training Program (eight 60-minute one-to-one treatment sessions delivered by a psychology graduate or teacher) ($n = 15$; $n = 14$ at follow-up)	No-treatment control ($n = 15$)	Treatment focus: interpersonal difficulties Outcomes measured: depression (MMPI), psychasthenia (MMPI) Measurement points: post treatment, 1.5 months post treatment

BDI, Beck Depression Inventory; BSI, Brief Symptom Inventory; DCP, Difficulties and Coping Profile Questionnaire; HADS, Hospital Anxiety and Depression Scale; HRSD, Hamilton Rating Scale for Depression; IES, Impact of Events Scale; LIFE, Longitudinal Interval Follow-up Evaluation; MAACL, Multiple Affect Adjective Checklist; MAS, Manifest Anxiety Scale; MSW, Master of Social Work; PSS-SR, Post-Traumatic Stress Disorder Symptom Scale Self-Report; RBPC, Revised Behaviour Problem Checklist; TAU, treatment as usual; TSCS, Tennessee Self-Concept Scale.

Setting and sample

The majority of the studies were conducted in North America (USA: $n = 7$;[59,61–66] Canada: $n = 1$[67]); the remaining studies were conducted in the UK.[56,60] Participants were recruited from and treatment was delivered in a variety of settings. The mean age of the sample was 15 or 16 years in most cases. Six of the 10 studies had an entirely male sample[56,59,61,62,65,67] and the sample was predominantly male in one other study;[66] only one study had an entirely female sample.[63] In the studies in which ethnicity was reported, the sample was predominantly Caucasian;[56,59,61,62,64,65] in one study the majority of the sample was African American.[66]

The total sample size was small for all of the studies. Only four studies had a total sample size > 50[61,62,64,65] and no studies had a total sample size > 100.

Interventions

The duration of the interventions ranged from six sessions lasting 60 minutes each[63] to 40 group sessions lasting 90 minutes each plus 20 one-to-one sessions lasting 60 minutes each (total 80 hours of treatment).[62] Five of the studies used a group format for treatment delivery,[59–61,64,65] three provided one-to-one treatment[56,63,67] and one used a combination.[62] A further study in which both the adolescent and the family were the focus of the intervention used a variety of forms of contact between the family and other professionals.[66] A range of professionals, typically with a professional mental health qualification, delivered the interventions, although one study did use doctoral students[63] and in another treatment was delivered by a psychologist and a doctoral student.[59] A brief summary of the different types of interventions used in the included studies is given below.

- *CBT.* Seven of the studies can be broadly classified as using some form of CBT.[56,59,61,63–65,67] CBT combines treatment strategies derived from cognitive and behavioural theories of the maintenance of emotional difficulties, such as anxiety and depression. Cognitive theories hypothesise that emotional difficulties are maintained because of cognitions or negative thoughts that are likely to be inaccurate or unhelpful in some way. These cognitions can in turn lead to behaviours that serve to maintain the emotional difficulty, because they prevent the person from learning that the thoughts are not accurate. Treatment strategies involve identifying the cognitions and seeking to test out the accuracy of those cognitions. The testing of these thoughts can take the form of verbal strategies (e.g. rating the evidence for and against a thought) or 'behavioural experiments' in which the person acts in a new way to test out the accuracy of the cognition.
  Behavioural treatment strategies are derived from classical conditioning and operant conditioning theories. Although these treatment strategies are based on behavioural rather than cognitive theories, in CBT they are often adapted to test out hypothesised key maintaining cognitions. Treatments based exclusively on behavioural principles are described separately.
- *Multisystemic Family Preservation Program.* Multisystemic family preservation treatment, used in one study,[66] draws on a range of therapeutic modalities including family therapy models, CBT and community consultation techniques. The goal of treatment is primarily to reduce delinquent behaviour and avoid out-of-home placements for the adolescent but, as part of this overall goal, the treatment recognises the need to meet other therapeutic objectives that may contribute to the primary goal, such as increasing positive family functioning. Unlike traditional psychological treatment, which is often office based and at a set time, in this approach therapists are on call 24 hours a day and meet with the family several times a week.
- *Problem-solving therapy.* One study used problem-solving therapy.[60] This treatment strategy involves breaking down problem-solving into a number of separate stages, teaching these stages to clients and then encouraging clients to apply these to real-life situations.[68] The original model identified five stages of problem-solving (general orientation, problem definition and formulation, generation of alternatives, decision-making and verification),[68] although other stages have been suggested. It was the original five-stage approach that was used in the study included here.
  The first formulation of problem-solving therapy used a behavioural framework[68] and the treatment is sometimes used as part of a CBT treatment, although it was originally intended to be, and can be, used as a stand-alone treatment.

- *Stress Management Training Program.* Stress management training was used in one of the studies.[67] It teaches progressive relaxation with the aim of reducing stress responses and is based on behavioural principles.
- *Other treatment modalities.* The study by Persons[62] described using psychotherapy as the intervention. The term 'psychotherapy' can be used as a generic term for any psychological treatment or can be used more specifically to refer to psychological interventions based on psychodynamic principles. Persons[62] does not state which theoretical principles formed the basis of the intervention, although the study does refer to the use of frequent interpretations, which typically indicates a psychodynamic approach. However, the study also refers to the use of negative reinforcement strategies, which suggests the use of behavioural principles. There is no reference to a treatment manual and so the exact content and theoretical basis of the intervention remains unclear.

Comparator

The majority of the studies used a two-arm design in which a single intervention was compared with some form of control condition, such as treatment as usual, wait-list control or no-treatment control. Two studies used a more complex design in which one or more treatment condition was compared with one or more control condition.[63,67] One further study compared two active interventions without some form of control condition.[61]

Quality assessment of the included studies

Table 11 summarises the risk of bias individually for the 10 included studies using the Cochrane risk of bias tool.[57]

TABLE 11 Quality assessment of the included clinical effectiveness studies

Study	Sequence generation	Allocation concealment	Blinding of participants	Blinding of outcome assessment	Incomplete data	Selective reporting	Other sources
Ahrens 2002[59]	Unclear	Unclear	High	High	Unclear	Low	Unclear
Biggam 2002[60]	Unclear	Unclear	High	Unclear	Low	Unclear	Unclear
Martsch 2005[61]	Unclear	Unclear	High	Unclear	High	High	Unclear
Mitchell 2011[56]	Low	Low	High	Low	Low	Low	Unclear
Persons 1966[62]	Unclear	Unclear	High	Unclear	Unclear	Unclear	Unclear
Reardon 1976[63]	Low	Unclear	High	Unclear	Unclear	Unclear	Unclear
Rohde 2004[64]	Low	High	High	Low	Low	Low	Unclear
Rohde 2004[65]	Low	High	High	Unclear	Unclear	Unclear	Unclear
Scherer 1994[66]	Unclear	Unclear	High	Unclear	High	Unclear	Unclear
Shivrattan 1988[67]	Unclear	Unclear	High	Unclear	Low	Low	Unclear

High, high risk of bias; low, low risk of bias; unclear, unclear risk of bias.

Sequence generation

Sequence generation assesses whether or not the process of assigning participants to the different arms of the trial was described in sufficient detail to assess if it is likely to produce comparable groups. Four studies were rated as being at low risk of bias for this item;[56,63–65] the remainder were rated as unclear. For many of the studies, therefore, it was unclear if the randomisation of participants to the trial arms was adequate to ensure that the groups were likely to be comparable at baseline. This increases the possibility that any observed differences in outcome post treatment may be due not to the intervention but to these pretreatment differences.

Allocation concealment

The allocation concealment item rates whether or not the allocation sequence could have been foreseen in advance by members of the research team and, therefore, subverted. There is evidence that inadequate allocation concealment is associated with inflated effect sizes relative to studies that had adequate concealment.[69] Only one study was rated as being at low risk for this item,[56] two studies were rated as high risk[64,65] and the remainder were rated as unclear. It is possible, then, that the observed effects of the interventions may be artificially inflated given that the majority of the studies were rated as having either an unclear risk of bias or a high risk of bias for this item.

Blinding

As expected, all of the studies were rated as being at high risk of bias for blinding of participants. For blinding of study outcome assessment, two studies were rated as being at low risk;[56,64] of the remaining studies one was rated as having a high risk of bias[59] and the remainder were rated as having an unclear risk of bias. Although the absence of blinding of participants is an inevitable risk of trials involving psychological treatments, blinding of outcome assessment can be achieved, although for most of the included studies it was unclear if attempts were made to ensure that personnel remained blind. Although in principle an absence of blinding could inflate or deflate an observed effect of treatment, the typical concern is that study personnel may seek to confirm the effectiveness of an intervention. This cannot therefore be ruled out for the majority of the clinical effectiveness studies considered here.

Incomplete outcome data

This item refers to the completeness of the outcome data for the main outcomes and includes exclusions from the analysis and attrition. Four studies were rated as being at low risk of bias,[56,60,64,67] two were rated as being at high risk of bias[61,66] and the remainder were rated as being at unclear risk of bias.

Selective outcome reporting

Selective outcome reporting occurs when researchers choose to report some outcomes but not others, with typically those measures showing more favourable outcomes being more likely to be reported. Four studies were rated as being at low risk of this form of bias,[56,59,64,67] one was rated as being at high risk[61] and the remainder were rated as being at unclear risk. The typical effect of selective reporting bias is to provide an artificially favourable view of the effectiveness of the intervention; this may be a problem for those six studies rated as being at high or unclear risk of bias.

Other sources of bias

All studies were rated as being at unclear risk of bias for the other potential sources of bias item. In general, the study method was often not well described, which led to the coding of unclear. For some studies there were additional methodological limitations that may also have acted as an additional source of bias. For example, Persons[62] measured baseline data after the allocation of participants, which may have altered participants' responses to the baseline measures.

Summary

In general, key methodological features of the trial design were poorly reported. Adequate reporting of these design characteristics is important; they determine the extent to which a study has sought to limit potential biases, many of which typically serve to inflate the observed effectiveness of the treatment. The generally poor reporting of these key methodological features means that it is possible that these biases may be present in many of the studies and that the observed effect of treatments may be inflated. It should be noted, however, that, although many of the items for many of the studies were rated as unclear, two studies were rated as being at low risk of bias across a number of domains.[56,64]

Effectiveness of the interventions

Mood disorders

Although a number of studies examined depression as an outcome and some targeted it as part of an intervention for a range of presenting difficulties, only one study examined an intervention specifically targeting depressive symptoms.[64] This study used the Adolescent Coping with Depression (CWD-A) course, an intervention that was initially developed and evaluated outside of forensic services. The sample consisted of adolescents who met criteria for major depression and conduct disorder. Participants were randomised to the intervention ($n = 45$) or a 'life skills' course ($n = 48$), which consisted of a current events review, training in life skills and academic tutoring. The authors do not state whether the life skills course was intended to be therapeutically effective for depression or an attention control condition. However, the content of the course does not appear to be linked to a recognised psychological model of depression and so it has been classified as being for a control condition. Both the intervention and the control condition were delivered in a group format. As described in the previous section, this was one of the two studies that were rated as being at low risk of bias across a number of the quality assessment domains.

The trial examined depression rates post treatment and at 6 and 12 months' follow-up using the Beck Depression Inventory-II (BDI-II),[70] the Hamilton Rating Scale for Depression (HRSD)[71] and diagnostic status (depressed vs. not depressed). The effect sizes are summarised in *Tables 12* and *13* (for relative risks, scores of < 1 favour the intervention; for SMDs, negative scores favour treatment). There were no differences between the two groups although the 95% CI for the HRSD post treatment was close to excluding zero with the effect in favour of the CBT group (SMD –0.39, 95% CI –0.81 to 0.02) (see *Table 13*). A diagnosis of conduct disorder was also reported as a secondary outcome. There were no differences between the two groups at any of the time points in terms of the diagnosis of conduct disorder (see *Table 12*).

TABLE 12 Depressive disorder clinical effectiveness trial: dichotomous outcomes

Study	Comparison	Outcomes	Time point	Relative risk (95% CI)
Rohde 2004[64]	CBT (group) vs. attention control (group)	Psychiatric diagnosis: depression	Post treatment	
			6 months post treatment	1.11 (0.65 to 1.90)
			12 months post treatment	0.99 (0.55 to 1.80)
		Conduct disorder	Post treatment	1.05 (0.76 to 1.45)
			6 months post treatment	1.01 (0.68 to 1.51)
			12 months post treatment	0.98 (0.63 to 1.50)

TABLE 13 Depressive disorder clinical effectiveness trial: continuous outcomes

Study	Comparison	Outcomes	Time point	SMD (95% CI)
Rohde 2004[64]	CBT (group) vs. attention control (group)	Depression:		
		(1) BDI-II	Post treatment	–0.17 (–0.59 to 0.24)
			6 months post treatment	0.04 (–0.39 to 0.46)
			12 months post treatment	0.26 (–0.16 to 0.68)
		(2) HRSD	Post treatment	–0.39 (–0.81 to 0.02)
			6 months post treatment	–0.03 (–0.45 to 0.39)
			12 months post treatment	0.26 (–0.16 to 0.68)

Anxiety disorders

As with the depression studies, although a number of trials examined anxiety as an outcome, only one study examined the effectiveness of an intervention specifically designed to target an anxiety presentation (*Table 14*).[59] Ahrens and Rexford[59] examined the effectiveness of group-delivered cognitive processing therapy for trauma symptoms ($n = 19$) compared with a wait-list control ($n = 19$). The treatment involved an educational phase in which participants were taught about the symptoms of PTSD, after which they conducted exercises to distinguish between cognitions and emotions, and evaluated beliefs and thoughts related to the traumatic event. Treatment also involved an exposure component in which participants were encouraged to produce a written or taped description of the traumatic event. Treatment was delivered using a group format.

The study used two measures of trauma symptoms [Impact of Events Scale (IES)[72] and Post-Traumatic Stress Disorder Symptom Scale Self-Report (PSS-SR)[73]] and also measured depressive symptoms as a secondary outcome using the BDI.[74] Measurement took place post treatment. As *Table 14* summarises, the study reported large and significant effects in favour of the intervention for both of the PTSD measures (PSS-SR: SMD –1.23, 95% CI –1.92 to –0.53; IES: SMD –1.53, 95% CI –2.26 to –0.80) and for the depression measure (BDI: SMD –1.44, 95% CI –2.15 to –0.72). Caution is needed, however, given that all of the risk of bias items assessed for this study were rated as either high or unclear except for one. It is possible, therefore, that the observed effect size is artificially inflated.

Disruptive disorders

We sought to identify trials that evaluated the effectiveness of interventions for a variety of disruptive disorders, including conduct disorder and ADHD. Although a number of trials looked at general aggressive or externalising outcomes, only two trials specifically evaluated the effectiveness of an intervention for a measure of conduct disorder and so met the inclusion criteria. No studies examining interventions for ADHD met the inclusion criteria.

TABLE 14 Anxiety disorders clinical effectiveness trial: continuous outcomes

Study	Comparison	Measure	Time point	SMD (95% CI)
Ahrens 2002[59]	CPT (group) vs. wait-list control	PTSD:		
		(1) PSS-SR	Post treatment	–1.23 (–1.92 to –0.53)
		(2) IES	Post treatment	–1.53 (–2.26 to –0.80)
		Depression: BDI	Post treatment	–1.44 (–2.15 to –0.72)

CPT, cognitive processing therapy; IES, Impact of Event Scale.

Martsch[61] compared two forms of group CBT as interventions for aggressive behaviour, including conduct disorder, as measured by the conduct disorder subscale of the Revised Behaviour Problem Checklist (RBPC).[75] The standard group CBT intervention taught anger management skills, social problem-solving and social skills. An additional two sessions, which parents were invited to attend, focused on parental–adolescent communication. The enhanced intervention covered the same CBT skills but provided an additional emphasis on group processes, including encouraging cohesion, participation by all members of the group, high levels of interaction and self-determination.

Measurement took place post treatment and at 9 months' follow-up. *Table 15* summarises the effect sizes for this study. The effect size post treatment was moderate in favour of the enhanced CBT group, although it fell short of conventional levels of statistical significant as indicated by the 95% CI (SMD –0.47, 95% CI –0.97 to 0.03). At 9 months' follow-up the two groups appeared broadly comparable (SMD –0.10, 95% CI –0.82 to 0.62). However, interpreting the follow-up data is difficult given the considerable attrition in both the standard CBT group (baseline $n = 31$, follow-up $n = 14$) and the enhanced CBT group (baseline $n = 34$, follow up $n = 17$). It should be noted that the original study reported results separately for older and younger boys and found that for the younger age group the standard intervention was more effective but for the older boys the enhanced intervention was more effective. Data from both groups were combined here to produce the effect sizes because it was not clear that there was an a priori rationale for assuming that the effects would be different for the two groups and that they should therefore be analysed separately.

Scherer *et al.*[66] examined a Multisystemic Family Preservation Program, which as described earlier involved multiple weekly meetings between the young person ($n = 23$) and his or her family members. This was compared with a treatment as usual condition ($n = 21$). Conduct disorder was measured using the RPBC,[75] as in the Martsch[61] study. Depression was also measured using the Brief Symptom Inventory (BSI).[76] Outcomes were measured post treatment. Effect sizes were not calculated for this study because of the substantial skew on both the measure of conduct disorder and the measure of depression. The authors analysed the data using analysis of variance (ANOVA) and report a significant time × condition interaction for the depression subscale [$F(1,41) = 6.12$, $p < 0.018$] but no significant differences for the measure of conduct disorder. However, ANOVA is a parametric test and, given the level of skew, considerable caution is needed in interpreting these results.

Treatments for other presentations

The remainder of the studies used an intervention that had a wider focus than a specific mental health presentation. Four examined interventions for increasing coping or reducing psychological distress[56,60,63,65] and two examined interventions for interpersonal difficulties.[62,67]

Summarising the results of these trials is difficult because the general focus of the interventions meant that often a very wide range of outcome measures was reported without the researchers clearly specifying a primary outcome measure. We have reported effect sizes for broad categories of outcomes that are likely to be improved by these general interventions (e.g. anxiety, depression).

TABLE 15 Disruptive disorders clinical effectiveness trial: continuous outcome

Study	Comparison	Outcome	Time point	SMD (95% CI)
Martsch 2005[61]	Enhanced CBT (group) vs. standard CBT (group)	Conduct disorder: RPBC	Post treatment	–0.47 (–0.97 to 0.03)
			9 months post treatment	–0.10 (–0.82 to 0.62)

Interventions for increasing coping/reducing psychological distress

Of the four studies that examined treatments to increase coping or reduce general distress, three used broadly similar interventions that could be classified as CBT[56,63,65] and one used problem-solving therapy,[60] which, as described earlier, can be incorporated into CBT. Despite this, it was it was not possible to conduct a meta-analysis as no two studies reported broadly similar outcomes measured at the same time point to permit a meaningful combination of data.

Three of the studies reported depression and anxiety outcomes.[56,60,63] Effect sizes could be calculated for only one study;[60] these are reported in *Table 16*. This study reported large and significant effects for depression post treatment (SMD –0.96, 95% CI –1.58 to –0.35) and at 3 months post treatment (SMD –1.01, 95% CI –1.62 to –0.39). Similar results were reported for anxiety post treatment (SMD –0.95, 95% CI –1.56 to –0.34) and at 3 months post treatment (SMD –0.82, 95% CI –1.42 to –0.21).

Mitchell *et al.*[56] compared CBT with treatment as usual. The data on the depression and anxiety outcomes for this study came from the Difficulties and Coping Profile Questionnaire (DCP), a measure developed for that study. The results showed substantial skew and so effect sizes were not calculated. The paper reported no significant effects for the depression and anxiety measures.

Reardon[63] compared two forms of CBT with an attention control and a no-treatment condition. It was not possible to calculate effect sizes for this study because of the limited reporting of the results. Reporting was restricted to means with no reporting of SDs and reporting of *p*-values limited to only those that were significant. The study used the depression and anxiety subscales of the Multiple Affect Adjective Checklist (MAACL) and analysed the results using ANOVA. It reported no significant effects for the anxiety subscale but a treatment × time effect for the depression subscale ($F = 3.220$, $p < 0.01$). Caution is needed in interpreting this result, however, given the level of skew.

Two studies[56,65] provided data on the internalising and externalising subscales of the YSR; the results are summarised in *Tables 17* and *18* respectively. Both studies compared CBT with treatment as usual. The effect sizes were small and none was significant.

TABLE 16 Clinical effectiveness trial of an intervention for increasing coping/reducing psychological distress: continuous outcomes – depression and anxiety

Study	Comparison	Outcome	Time point	SMD (95% CI)
Biggam 2002[60]	Problem-solving therapy vs. TAU	Depression: HADS	Post treatment	–0.96 (–1.58 to –0.35)
			3 months post treatment	–1.01 (–1.62 to –0.39)
		Anxiety: HADS	Post treatment	–0.95 (–1.56 to –0.34)
			3 months post treatment	–0.82 (–1.42 to –0.21)

HADS, Hospital Anxiety and Depression Scale; TAU, treatment as usual.

TABLE 17 Clinical effectiveness trials of interventions for increasing coping/reducing psychological distress: continuous outcome – internalising symptoms

Study	Comparison	Outcome	Time point	SMD (95% CI)
Mitchell 2011[56]	CBT vs. TAU	Internalising: YSR	11 months post treatment	0.20 (–0.44 to 0.84)
Rohde 2004[65]	CBT vs. TAU	Internalising: YSR	Post treatment	–0.04 (–0.53 to 0.44)

TAU, treatment as usual.

TABLE 18 Clinical effectiveness trials of interventions for increasing coping/reducing psychological distress: continuous outcome – externalising symptoms

Study	Comparison	Outcome	Time point	SMD (95% CI)
Mitchell 2011[56]	CBT vs. TAU	Externalising: YSR	11 months post treatment	–0.40 (–1.04 to 0.25)
Rohde 2004[65]	CBT vs. TAU	Externalising: YSR	Post treatment	–0.05 (–0.53 to 0.44)

TAU, treatment as usual.

Interventions for interpersonal functioning

Two of the older studies examined interventions designed to improve interpersonal functioning.[62,67] Despite the stated focus of these studies, neither examined social functioning as an outcome measure; instead, both measured a range of mental health symptoms using among other measures the MMPI.

Persons[62] examined a psychotherapy intervention compared with treatment as usual. As described earlier, the exact content of the intervention and its theoretical foundation were not well described. *Table 19* summarises the results of this study for the depression and anxiety outcomes (for this study, calculations of effect sizes are based on SDs estimated from reported *p*-values). The effect sizes suggest no difference between the two groups on the depression measure, although the effect size for one of the two anxiety measures, the Manifest Anxiety Scale (MAS), suggests that the intervention was more effective than treatment as usual.

Shivrattan[67] compared a CBT intervention with a focus on social interaction skills, a stress management training programme and a no-treatment control condition. Effect sizes were not calculated for this study because reporting was restricted to means with no reporting of SDs and reporting of *p*-values was limited to only those that were significant. The study reported no differences on the MMPI subscales that linked to the mental health outcomes of interest for the current review, such as depression or psychanthesia (excessive doubts, compulsions, obsession and unreasonable fears).

Summary

There were too few studies to make firm conclusions about the clinical effectiveness of interventions for mental health problems in young people who offend. No conclusions can be made about the effectiveness of psychopharmacological interventions because no studies were identified that met our inclusion criteria. Any conclusions about psychological interventions are also tentative given the limited number of studies identified. Those studies that did meet the inclusion criteria tended to be small and many had methodological problems or key methodological features were poorly reported. There was an insufficient number of studies using broadly comparable interventions for broadly comparable psychological difficulties to conduct a meta-analysis. This would have compensated for the small size of many of the studies and would have allowed an assessment of the extent to which variations in effect sizes were associated in a systematic way with variations in the methodological quality of the primary studies.

TABLE 19 Interventions for interpersonal functioning: continuous outcomes (depression and anxiety)

Study	Comparison	Outcome	Time point	SMD (95% CI)
Persons 1966[62]	Psychotherapy vs. TAU	Depression: MMPI subscale	Post treatment	–0.75 (–1.20 to 0.31)
		Anxiety:		
		MAS	Post treatment	–0.64 (–1.08 to –0.19)
		MMPI subscale	Post treatment	–0.75 (–1.20 to 0.31)

Many of the interventions used in the studies had a broad focus and a clearly identified primary outcome was not specified. Instead, results for a range of outcomes were reported. Clear specification of the primary outcome is important to protect against post hoc selective reporting and may be particularly important for studies in this area with a general treatment focus, such as improving coping, for which there may not be an obvious primary outcome.

Other methodological limitations or poor reporting of methodological features were also present in a number of studies (e.g. lack of allocation concealment, lack of blinding of outcome assessment). The general effect of these would be to artificially inflate the observed clinical effectiveness. Although some studies did report broadly encouraging findings, caution is therefore needed. This conclusion is in keeping with that of an early systematic review in this area.[18] The current and previous reviews used somewhat different inclusion criteria. Each review identified 10 studies, of which six were common to both. Although Townsend *et al.*[18] performed a meta-analysis of three group CBT studies that had depression as an outcome, we chose not to conduct a meta-analysis. The meta-analysis of Townsend *et al.*[18] combined studies that had depression as a primary treatment focus[64] with other studies that had a distinct focus, such as the Ahrens and Rexford[59] trial in which the treatment focus was for PTSD symptoms. We judged that it was not meaningful to combine studies in which depression was a primary focus with those in which it was a secondary outcome.

One of the objectives of the review was to identify in which groups it may be of use to screen and offer interventions. There were too few studies to establish the presentations for which screening and offering interventions may be of use and too few studies of both community and incarcerated settings to make conclusions in terms of these populations. However, the review did identify trials of depression, PTSD and conduct disorder. Studies of diagnostic test accuracy were also identified for these three presentations. This suggests that these may be potential candidates for the development of an exemplar decision model, as discussed in *Chapter 8*. The clinical effectiveness data, however, suggest that depression may be the most suitable exemplar. The one depression trial,[64] a study of a CBT package for depression, was one of only two studies to be rated as being at low risk of bias across a number of the quality assessment domains. In contrast, the anxiety study[59] was rated as being at either unclear or high risk of bias across a number of domains. Of the two studies examining conduct disorder as an outcome, one[61] was rated as being at high or unclear risk of bias across all quality assessment domains. In addition, this two-arm trial compared two forms of CBT against each; it did not include a usual care comparator, which would be of use in the modelling phase. For the second study it was not possible to extract relevant estimates of effect sizes.[66] As will be discussed in more detail in *Chapter 8*, any decision model based on the available data will be at best an illustrative example of the type of modelling that could be undertaken in the area in future when a larger number of higher-quality studies is available.

Reflections on policy and practice

As with the results for the diagnostic test accuracy studies, there may be concerns about the extent to which the findings on clinical effectiveness can be generalised to a UK setting. Many of the studies were conduced in the USA and a number were conduced in incarcerated settings. As described in *Chapter 1*, the majority of young people who offend in the UK are managed in the community. Two UK studies were included in the review; however, these were both conducted in custodial settings.

Although screening is currently recommended for the identification of mental health problems in young people who offend, screening strategies assume that there are clinically effective treatments that can be offered to people who are identified as requiring an intervention. Although some positive results were reported in the identified trials, there remains substantial uncertainty around the clinical effectiveness of interventions for mental health problems in this population.

Chapter 7 Cost-effectiveness of methods to identify and treat mental health difficulties in young people who have offended

The previous systematic reviews in this report have identified some, albeit limited, evidence on diagnostic test accuracy and clinical effectiveness. The next stage of the review sought to identify relevant evidence on the cost-effectiveness of identification strategies and interventions for mental health problems among young people who have offended.

Methods

In this phase of the review we sought to answer the following questions:

- What is the cost-effectiveness of screening strategies for mental health difficulties among young people who have offended?
- What is the cost-effectiveness of interventions for mental health difficulties in young people who have offended?

The search strategy outlined in *Chapter 3*, which included searches of NHS EED and HEED, was used to identify relevant studies.

Inclusion and exclusion criteria

Articles were eligible for inclusion if they were full economic evaluations of identification strategies or clinical interventions for mental health problems among young people who have offended. Full economic evaluations refer to an assessment of costs and outcomes of any identification or intervention strategy against those of an alternative (cost–benefit analyses, cost-effectiveness analyses, cost–utility analyses).

The PICO criteria that guided study selection were:

- *Population and setting:* young people (aged 10–21 years) who have offended and who are in contact with the criminal justice system.
- *Intervention:* any screening strategy or any psychological or pharmacological intervention for the specified list of mental health problems.
- *Comparison:* for studies reporting the cost-effectiveness of a screening strategy, the comparator will be usual care; for studies reporting the cost-effectiveness of an intervention, the comparator will be no intervention, placebo, attention control or other 'psychological placebo' or usual care.
- *Outcome:* any measurement of cost.
- *Study design:* economic evaluations (cost–benefit analyses, cost-effectiveness analyses, cost–utility analyses) of screening strategies or interventions.

Data extraction

Two reviewers screened titles and abstracts to identify potentially eligible studies. Any disagreements were resolved by consensus or deferred to a third party if necessary. Full papers of potentially eligible studies were obtained and assessed for inclusion independently by two reviewers. Our initial scoping of the literature had indicated that the cost-effectiveness literature in this area was likely to be limited; therefore, liberal criteria were applied at both stages and any potentially relevant papers were passed on to one of the health economists for further evaluation.

This evaluation of evidence followed explicit guidelines laid down by the Centre for Reviews and Dissemination in the preparation of the NHS EED.[77] The quality and relevance of any available economic data were judged from the perspective of the UK NHS according to criteria laid down by Drummond.[78]

Results

No studies were identified that met the inclusion criteria for this stage of the review.

Summary

Despite significant discussion about and policy recommendations on strategies to detect mental health needs among young people who offend, the review identified no studies that were full economic evaluations of identification strategies or clinical interventions for mental health problems among young people who have offended.

Reflections on policy and practice

Although there are current policy guidelines recommending screening and on the basis of this intervening for mental health problems in young people who offend, no studies addressing the cost-effectiveness of these policies were identified.

Chapter 8 Decision model

Mental health conditions are highly prevalent in young offenders[9] but, given the limited evidence on diagnostic accuracy and clinical effectiveness, a decision was made to constrain the policy question addressed by the decision model to focus on the screening and subsequent management of one common unmet need in the young offending population: depression.

The rationale for constraining the policy question and developing an 'exemplar' case study for the decision model was based on the following reasons: (1) depression is highly prevalent in young offenders (up to 15% within the UK system); (2) as discussed in *Chapters 4* and *6*, taken together there is more evidence on the effectiveness of screening for depression and treatment in young offenders than on the effectiveness of screening for other mental health conditions; and (3) unlike other common mental health problems found in young offenders (e.g. conduct disorder), depression is not externalising and may therefore be more likely to go undetected.

The findings of the decision model should be considered within the limitations of the available evidence emerging from the systematic reviews of diagnostic accuracy and clinical effectiveness studies. However, the decision model makes an important contribution to the overall evidence by providing an exemplar based on a formal quantitative framework that provides a clear indication of the various inputs and data sources required to appropriately inform cost-effectiveness assessments. Importantly, the model provides an iterative basis for updating and revisiting the findings as new evidence emerges in the future.

Setting the decision context

Detection and treatment of mental health conditions in young offenders is, by definition, an intersectoral issue. When developing a decision-analytic model it is important to establish the context for informing resource allocation decisions.

The decision analysis primarily considers costs and health outcomes [expressed in quality-adjusted life-years (QALYs)] from the perspective of the UK health services. Current policy stipulates that screening occurs at the initial contact with the criminal justice system or at the first available opportunity, whereas the treatment pathway is expected to follow based on the outcome of screening.

Adopting this conventional health service-only perspective could infer certain limitations as the costs and benefits of treating mental health issues in young offenders may go beyond the health-care system. To capture further intersectoral effects, supplemental analysis extends the perspective to consider costs and benefits that may be realised by the youth justice system in terms of future crimes averted as a result of treating depression in young offenders.

The decision problem

Joint initiatives between the Youth Justice Board and the Department of Health aim to implement the CHAT, which in part assesses mental health. To date, UK policy has provided guidance on screening for mental health problems in young people who offend;[79] however, the clinical and economic benefits of this policy remain to be demonstrated. Providing this 'exemplar' decision model for depression represents an important first step in bringing together available evidence to develop a framework for future decision-making.

Ideally, specification of a decision model should facilitate comparisons of all identification strategies that could feasibly be used in the NHS and/or youth justice system. The systematic review of all available evidence provided a range of potential parameters for identification and treatment for modelling; however, there remain several unknowns required to appropriately address the decision problem. The evidence that

is currently available (as reviewed in previous chapters) places important constraints on the structural framework of any potential model and this exemplar provides a preliminary basis to inform the future research agenda within the context of current decision uncertainty.

A further constraint within the context of the decision problem is specifying the appropriate perspective to address the intersectoral nature of the problem. For example, health-care provision within the youth justice system raises additional complexities for decision-makers over how the benefits of identification and treatment are to be valued for decision-making. Furthermore, as unaddressed mental health need can increase criminal justice costs, the question is how to value benefits not directly relevant to the conventional health-care decision-maker (such as a reduction in reoffending rates after treating depression). The model provides a basis for considering where the main costs and benefits are being incurred and areas where broader improvements in public sector efficiency may be possible.

Methods

The objective of the exemplar model was to estimate, based on best available data, the costs and health outcomes for a range of feasible identification strategies. As already explained, the analysis is primarily set to the conventional health services perspective and this initial perspective is extended to also consider the youth justice system perspective. All costs are expressed in present-day values (2013) and health outcomes are expressed in QALYs. The time horizon for the analysis of depression is 1 year; hence, no discounting of costs or benefits was applied.

The model was made up of two parts: (1) an identification model, reflecting the diagnostic performance and administration costs of the alternative identification strategies, and (2) a treatment model that evaluated the subsequent costs and outcomes (expressed in QALYs). To consider intersectoral implications, the effect of treatment on recidivism rates was incorporated as a cost offset against the cost of the identification strategy.

For an identification strategy to be cost-effective, it is important that a cost-effective treatment strategy is available. Without a known cost-effective treatment, identification of an effective screening strategy would not be useful for the decision-maker because the identified patients would not be offered a cost-effective treatment. Hence, a treatment model is required to evaluate the cost-effectiveness of identification strategies in terms of the health benefits of identifying mental health conditions. Given the importance of the subsequent treatment, the treatment model for young offenders with depression is considered first.

Table 20 summarises the stages of the analysis, detailing sequentially the screening strategies evaluated and the perspectives adopted.

TABLE 20 Summary of the stages of analysis

Screening strategy	Perspective for the cost-effectiveness analysis
Single-stage screening	Health services
Two-stage screening	Health services and intersectoral

The treatment model

To allow decision-makers to evaluate cost-effectiveness evidence across health conditions, generic assessments of health outcomes should be expressed in terms of QALYs. The systematic review of cost-effectiveness evidence (see *Chapter 7*) identified no previous studies within the young offender population to inform this treatment model. As such, a bespoke treatment model was developed to serve as the exemplar for the management of depression in young offenders.

The systematic review of effectiveness studies of depression in young offenders identified one study that provided relevant evidence on the effectiveness of treating depression in young offenders.[64] This study reported the health outcome in terms of depression-free days (DFDs), which is not a generic measure of health and would provide limited evidence to allow decision-makers to compare cost-effectiveness evidence across health conditions. Hence, to construct the treatment model in the absence of generic measurements of health, a mapping approach was used to translate DFDs into generic health-related utility measures by assigning a utility value to each DFD,[80–82] which then allowed us to calculate QALYs over the period of the study.

As this model aims to serve primarily as an exemplar, in the modelling strategy we opted to avoid the additional complexities of developing a de novo disease-specific model and selected from the identified literature on clinical effectiveness any study containing the prerequisites to model QALYs directly.

To reflect uncertainty in the input values, the sensitivity of the input parameters was evaluated using deterministic sensitivity analysis.

Parameter inputs for the treatment model

The intervention

Rohde *et al.*[64] evaluated the clinical effectiveness of group CBT compared with a life skills control condition within a young offending population with major depressive disorder. As described in more detail in *Chapter 6*, the group CBT programme used the CWD-A course, which was provided to an average group size of 10.4 participants. The control condition was a life skills course in which young offenders reviewed recent events and received life skills training (such as filling out job applications) and academic tutoring. Although usual care within the UK youth justice system is unclear, life skills were generally consistent with the expected usual care arrangements for young offenders.

Relative treatment effect

The major depressive disorder recovery rate post treatment was 39% for the CWD-A course and 19% for life skills. Furthermore, over 64 weeks a Kaplan–Meier product-limit survival curve provides the proportion of individuals recovering from the depressed state at each 4-week interval. Extracting these data on recovery over the period of the study, DFDs were calculated as the number of days that the average individual in each group was depressed. Health-related utility values were assigned to the DFDs and days depressed for each month.[81] These were summed over the study period using an area under the curve approach and QALYs were calculated (see *Appendix 6* for further details).

Depression and recidivism

The evidence in the literature suggests that depressed young offenders are more likely to reoffend than non-depressed young offenders. To incorporate the impact of treating depression on the recidivism rate, Harshbarger[83] estimated the relative risk of reoffending given depression (compared with no depression) as 1.3034 per year.

Overall, Ministry of Justice statistics[84] report that 35.3% of offenders reoffend per year. Taking this as the overall probability of reoffending [i.e. $P(\text{Reoffend}) = 0.353$] and adjusting the probability based on data from Harshbarger,[83] the probability of reoffending conditional on being depressed [i.e. $P(\text{ReoffendIDepression})$] was estimated to be 0.441.

Impact of cognitive–behavioural therapy on recidivism

In a Campbell review, Lipsey *et al.*[85] report the effect of CBT on recidivism (using the criterion of no further offending in the 12 months after the intervention). The meta-analysis includes 58 studies of the effectiveness of the intervention as a mean odds ratio (i.e. the odds of *not* reoffending in the subsequent 12 months following treatment compared with the control). The mean odds ratio was 1.53 ($p < 0.001$), which implies that offenders receiving CBT were one and a half times more likely to *not* reoffend within 12 months post treatment than those not receiving CBT.

Incorporating the mean odds ratio for the effect of CBT on recidivism into the conditional probability of reoffending given being depressed, the probability of reoffending given being depressed and having received CBT was derived as 0.34. This conditional probability is utilised to estimate the expected reduction in recidivism for individuals with depression receiving CBT – this approach also highlights the potential importance of the intersectoral perspective. Assigning this probability of reoffending assumes that CBT for depression has the same impact on recidivism as CBT treatments that may more directly target recidivism, which may not be the case. This uncertainty is explored in the sensitivity analyses.

Resource utilisation and cost inputs

Cost of group cognitive–behavioural therapy The estimates of unit costs relevant to the health service were taken from Netten and Curtis[86] and specific costs relevant to the justice system were taken from Brookes *et al.*[87] These were combined with the intervention protocol described in the Rohde *et al.*[64] study. The intervention in the Rohde study was made up of 16 sessions each lasting for 2 hours and the average group size for the programme was 10.4 participants. This information was used to estimate the cost of CBT per participant. Rohde *et al.*[64] state that the intervention included 'two interventionists to better monitor in-session behaviour' (p. 662); however, as it is unclear if this potential modification of treatment delivery (i.e. two therapists per session) may apply to the UK criminal justice system, the cost-effectiveness of treatment is calculated for both one and two interventionists.

Cost of crime averted The effectiveness of CBT in reducing rates of recidivism was translated into the expected number of crimes averted. The Ministry of Justice[84] reports the reoffending of juveniles during 2010–11. Taking the percentage change in number of offenders since 2000 (2.5%), the expected numbers of offenders and reoffenders were estimated for 2013.

Home Office research study 217[88] reports estimates of the costs of crime under 'notifiable offence categories'. Costs per category are divided into three categories to reflect the true costs of crime: the anticipation of crimes (costs of security and insurance); the consequence of criminal events (e.g. value stolen and damaged, emotional and physical impacts, and impacts on health services); and responses to crime (costs spent tackling criminals to the criminal justice system).

In the model, Ministry of Justice data provide expectations around offending and reoffending and Home Office data provide estimates of the monetary value of potential reductions in crime (for more specific details see *Appendix 6*).

Consumption value of health benefits on crime

The National Institute for Health and Care Excellence (NICE) uses cost-effectiveness analysis as a means of comparing the expected health benefits for additional costs falling on the fixed health-care budget. For the case presented within this review, economic effects that occur outside the health-care system would require a wider 'societal perspective'. This raises issues of the relevance of non-health benefits to the restricted NHS budget. To accommodate the wider perspective required when considering mental health problems in youth justice settings, the value of the benefits occurring outside of the remit of the health system (i.e. within the wider society) must be explored. However, to ensure that any analyses remain relevant to the restricted health budget, alternative policy perspectives may be adjusted by applying the notion of the consumption value of health, that is, the amount of consumption that is equivalent to 1 unit of health (see Claxton *et al.*[89] for a more comprehensive summary of the approach).

Extending the perspective conventionally adopted by NICE to consider the net consumption value of health raises empirical questions. The consumption value approach was taken to integrate costs of crime averted into the economic evaluation by assuming a consumption value of health of £60,000. This reflects the fact that costs and benefits falling outside of the health system may not be valued the same by a health services decision-maker or, if they are, the willingness-to-pay (WTP) threshold of the health services decision-maker is likely to be higher than the conventional threshold used for economic evaluation from a health services perspective. Decision-makers' WTP (given the level of uncertainty in the parameters) is likely to be closer to the lower bound of £20,000 per QALY. As such, the incorporation of the cost of crime offset through treatment was down-weighted by a factor of three.

Utility weights: converting depression-free days to quality-adjusted life-years Depression-free days extrapolated from the Rohde *et al.*[64] study indicate the incremental number of days per individual without depression. Revicki and Wood[81] provide health-related utility weights, which can be applied to the mild (0.685), moderate (0.59) and non-depressed (0.85) states. DFDs indicate the proportion of total time spent in non-depressed and depressed states; they provide the basis for weighting using the identified utility weights for depression.[81]

Rohde *et al.*[64] report at baseline an average score on the BDI of 16.6 for a cohort of non-incarcerated adolescents with comorbid major depression and conduct disorder; as this would indicate that the majority of the cohort had mild depression,[90] a Revicki and Wood[81] utility weight for mild depression (0.685) was assumed for the whole population (as described subsequently, this assumption was subjected to sensitivity analysis). The non-depressed state was weighted by 0.85.

Health utilities were calculated for the full study period (64 weeks) for both group CBT and the control condition. Incremental QALYs of treatment are the differences between the two groups averaged over 52 weeks (for further details see *Appendix 6*).

Implications of the treatment model (health and intersectoral)

Before considering whether or not identification strategies represent good value for money, it is worth reiterating that a cost-effective treatment should be identified first. Although the Rohde *et al.*[64] study provides the means of constructing a model (i.e. by estimating DFDs and therefore QALYs), it should be noted upfront that the study included only 93 adolescents and had low power to detect a statistically significant difference between the groups. More importantly, this model highlights the need for larger and more definitive clinical trials to better inform future decision-making.

Estimating incremental QALYs from treating young offenders with depression using a group CBT approach suggests that an individual would gain 0.0113 QALYs compared with the control condition. The cost of the 16 group sessions in the CBT programme is calculated to be £2054 with one interventionist or £3910 assuming the modified treatment protocol in which two therapists were used. Per individual, the average cost would be £197.51 and £375.97 respectively. Adopting primarily the health-service perspective on treatment, this suggests that in the best-case scenario group CBT would cost £17,542 per QALY and using the modified protocol (two therapists per session) group CBT would cost £33,393 per QALY.

Applying NICE's WTP threshold of £20,000–30,000 would suggest that only single-therapist group CBT is cost-effective, with the modified protocol not representing value for money. As such, evaluation within the identification model uses the conventional single interventionist to provide group CBT in the youth justice setting.

The identification model

The identification model assumed a decision tree structure to capture the outcomes of varying the sensitivity and specificity of the diagnostic strategies. Four possible outcomes are considered for each diagnostic strategy and relevant costs and outcomes evaluated for: (1) true positive, (2) false negative,

(3) true negative and (4) false positive. The identification model is driven by the prevalence of depression, the sensitivity and specificity of specific diagnostic tools and the cost associated with each strategy.

Strategies evaluated

The decision problem would ideally compare *all* potential identification strategies that are feasible to implement within the youth justice system. However, in reality, evaluation of these options has been constrained by the availability of evidence. *Chapter 4* was used to inform the identification strategies considered in the economic analysis and only studies containing sufficient data were used to form the basis of the parameters included.

The base-case analysis considered single-stage screening followed by treatment or no treatment for depression based on the outcome of screening. In addition to the base-case analysis, separate scenarios were considered that explored a range of alternative strategies (discussed in more detail subsequently). The alternative approaches considered included (1) the effects of two-stage screening in which the second stage uses a gold standard confirmatory interview; (2) the impact of diagnostic accuracy on recidivism; and (3) varying estimates of input parameters in the model (such as prevalence).

The diagnostic component

Model structure and key assumptions

To consider the implications of screening populations entering the youth justice system and the potential gains for those specifically with depression, *Figure 4* illustrates the outcomes of screening and treatment and describes the associated costs and outcomes considered.

Within this screening framework, all young offenders (depressed or not) entering the criminal justice system are screened on first contact with the system (as per guidelines under CHAT). For the purpose of the current analysis it is assumed that mental health need will be unknown up until the point of screening.

Each individual entering the system would receive an intervention strategy incurring the cost of the related screen. Here we consider the time taken by the health-care professional (within the criminal justice system) to conduct the screening as the main cost element.

The Offender Health Research Network indicates that reception health screening is typically carried out by nurses.[91] To apply the specific unit cost to the required time for screening, a value of £24 per hour for 'Prisons: Nurse (mental health)' taken from Brookes *et al.*[87] was used. The specific cost parameter for each individual diagnostic tool evaluated is presented in *Table 21*.

Of the total screened, those expected to screen positive (either true positive or false positive) receive the intervention for depression (e.g. group CBT). Dependent on the distribution across the four diagnostic outcomes, the expected number of QALYs and the cost were calculated. For individuals who screened as false negative (i.e. screened negative despite depression), the mental health need is unmet and the risk of reoffending remains high.

To inform the expected prevalence of depression within the youth justice system, we used data from Fazel *et al.*,[2] who determined prevalence rates of mental disorders (including depression) among adolescents in juvenile detention using a summary of 25 surveys. The results from this review were extracted and presented by gender, indicating that the depression rate in males is 11% (95% CI 7% to 14%) and that in females is 29% (95% CI 22% to 37%).

Sensitivity analysis

Sensitivity analysis explored the impact of considering alternative values for three key drivers in the model: (1) prevalence of depression under usual care; (2) utility weights assigned to DFDs; and (3) the effect of CBT on recidivism.

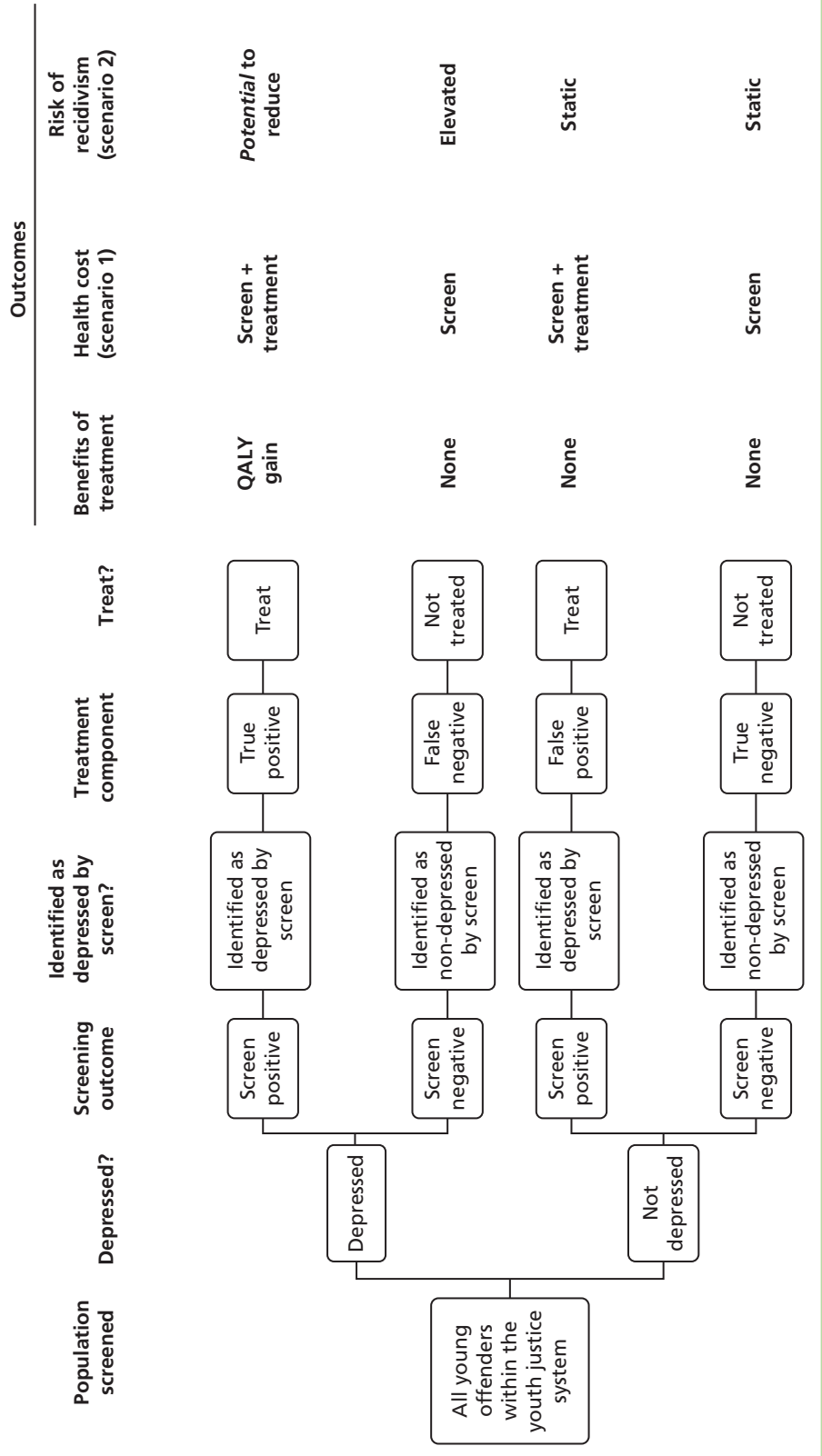


FIGURE 4 Schematic of the identification model and the implied treatment outcomes.

TABLE 21 Screening tools, administration time and associated costs

Screening tool	Administration time (minutes)	Cost (£)	Source
MAYSI-2 (paper)	10	4	Wasserman *et al.*[43]
MMPI-A	90	36	Cashel *et al.*[35]
MAYSI-2 (paper)	10	4	Hayes *et al.*[38]
MAYSI-2 (voice)	10	4	
DPS	15	6	McReynolds *et al.*[41]
MFQ	6	2.4	Kuo *et al.*[40]
Short MFQ	2.5	1	
MAYSI-2 (paper)	10	4	

Results

The results are presented in two parts: (1) the primary (base-case) results including the outputs of the base-case model (as outlined above) and (2) sensitivity analysis evaluating the impact of varying input parameters on the cost-effectiveness analysis. Both sets of results are presented using two perspectives, that is, the health services perspective and an intersectoral perspective. The intersectoral analysis estimates the cost offset to the youth justice system through reduced criminal activity as a result of identifying and treating depressed offenders.

Primary results

The analysis was conducted for single-stage screening and two-stage screening strategies. Single-stage screening may use the gold standard approach (i.e. DISC with a sensitivity and specificity of 1) or may use a relatively imperfect but less resource-intensive screening tool with a sensitivity and specificity of < 1 (such as MAYSI-2). Single-stage screening is followed by a treatment decision based on the outcome of screening. The two-stage screening strategy involves the administration of a gold standard instrument on individuals who screened positive (both true and false positives) on the single-stage screening measure (as expected, two-stage screening does not apply to the case when the gold standard tool is used as the first screening tool).

Cost-effectiveness of using single-stage screening tools from a health-care perspective

Table 22 presents the cost-effectiveness estimates of single-stage screening compared with usual practice. Usual practice for the purposes of the model is initially simulated as no active detection (i.e. no formal screening employed). Note that, in terms of treatment costs, the primary analysis assumed that one therapist would be present during the group CBT sessions.

In general, *Table 22* shows that, compared with current practice (i.e. no active screening), single-screening strategies identified in the systematic review are not likely to be cost-effective based on the commonly used WTP threshold in the UK for an additional unit of health outcome (i.e. £20,000–30,000 per QALY).

As emphasised earlier, there is limited evidence on the effectiveness and cost-effectiveness of treatments for depression in young offenders, which suggests that, even if screening had no cost and had a sensitivity and specificity of 1 (which is unrealistic), treating the true positives with group CBT would cost £17,542 per QALY (based on a health services perspective). Therefore, introducing screening costs would only increase the cost per QALY because of the additional costs of screening and the loss of false-negative cases who would have benefited from treatment. Hence, it is not surprising that none of the single screening strategies was found to be cost-effective. This further reinforces that, for screening to be cost-effective, it is important to have a treatment strategy that produces a significant improvement in outcomes at a relatively low cost.

TABLE 22 Cost-effectiveness of single-stage detection strategies to inform the treatment decision (health-care perspective)

		MAYSI-2							
	Gold standard	Paper version	Paper version	Voice version	Paper version	MFQ	Short MFQ	DPS	MMPI-A
Study (no. of screening tests in the study)	e.g. DISC (NA)	Wasserman 2004[43]	Hayes 2005[38] (1)	Hayes 2005[38] (2)	Kuo 2005[40] (3)	Kuo 2005[40] (1)	Kuo 2005[40] (2)	McReynolds 2007[41]	Cashel 1998[35]
Cost per QALY (single screening tools) (£)	43,195	77,378	83,156	61,844	71,804	42,839	52,174	52,322	78,662

NA, not applicable.

Cost-effectiveness of two-stage screening strategies from a health-care perspective

Use of single screening tools alone is not cost-effective as the imperfect accuracy of each tool implies that false positives receive unnecessary treatment, increasing the overall cost, and false negatives miss out on the treatment and potential gains in outcome. Although the second stage (the gold standard) in the two-stage screening process does not deal with the issue of false negatives, it differentiates the true positives from the false positives, ensuring that treatment is provided only to those who are truly depressed.

Table 23 presents the comparison between this two-stage strategy and usual practice. The cost per QALY analysis suggests a significant efficiency gain over the use of single screening tools alone. However, only the MFQ and Short MFQ (disease-specific tools for depression) are potentially within the upper bound of the WTP threshold of £20,000–30,000 per QALY.

It should be noted, however, that the analyses of both single and two-stage screening assume that there is no cost associated with false negatives. This assumption may not hold in practice; however, we did not find any evidence on health service resource use because of untreated depression in the young offending population.

Cost-effectiveness of two-stage screening strategies from an intersectoral perspective

The decision to treat a mental health need in young offenders may have wider-reaching benefits than the health outcomes alone. In adopting the intersectoral perspective of the health and youth justice services, the aim is to examine the extent of the potential costs and benefits of various detection strategies; this would be relevant if an improved mental health status had consequences for future criminal behaviour (specifically, changes in recidivism rates).

Table 24 presents the results of the cost-effectiveness analysis including the expected cost offset from the potential reductions in recidivism rates. Given that only the two-stage detection strategies were found to be cost-effective within the health-care perspective, only these strategies were analysed from an intersectoral perspective.

Including the expected cost offset from reduced recidivism to the evaluated two-stage detection strategies, all strategies except for MMPI-A fall within the range of the decision-maker's WTP. This model provides a provisional indication that gains in health status (through appropriate detection and treatment) may be justified by the potential cost offset to the youth justice system. However, it should be noted that the base-case intersectoral analysis gives equal weight to costs incurred or saved by the health and youth justice systems (i.e. the analysis assumes that costs incurred by the health system can be compensated in a 3 : 1 ratio by cost savings in the youth justice system).

To compare strategies and indicate which may represent the most efficient use of resources, results can be presented on a cost-effectiveness frontier. The frontier connects incremental cost-effectiveness ratios (ICERs) of strategies on the cost-effectiveness plane to identify strategies that dominate other less cost-effective strategies. Strategies that lie on the frontier line represent value for money and options falling on the line are compared to indicate whether or not incremental health benefits justify any incremental costs. The incremental analysis compared with no active detection is presented in the cost-effectiveness plane in *Figure 5.*

Examining the cost-effectiveness plane suggests that two-stage screening using the MFQ or the Short MFQ in the first stage is most cost-effective, as represented by the cost-effectiveness frontier on the cost-effectiveness plane. The cost-effectiveness frontier represents the most efficient points among all screening strategies examined (i.e. the frontier represents the strategies that for a given level of effect have the highest cost or vice versa).

TABLE 23 Cost-effectiveness of two-stage screening strategies to inform the treatment decision (health-care perspective)

		MAYSI-2							
	Gold standard[a]	Paper version	Paper version	Voice version	Paper version	MFQ	Short MFQ	DPS	MMPI-A
Study (no. of screening tests in the study)	e.g. DISC (NA)	Wasserman 2004[43]	Hayes 2005[38] (1)	Hayes 2005[38] (2)	Kuo 2005[40] (3)	Kuo 2005[40] (1)	Kuo 2005[40] (2)	McReynolds 2007[41]	Cashel 1998[35]
Cost per QALY (two-stage strategy) (£)	43,195	36,632	38,185	33,329	35,521	28,278	29,118	33,915	55,910

NA, not applicable.
a Here, the gold standard (e.g. DISC) is a single-stage screen and is included in the table to provide the reference standard.

TABLE 24 Cost-effectiveness of two-stage screening strategies with the inclusion of the cost offset from reduced rates of recidivism (intersectoral perspective)

		MAYSI-2							
	Gold standard[a]	Paper	Paper	Voice	Paper	MFQ	Short MFQ	DPS	MMPI-A
Study (no. of screening tests in the study)	e.g. DISC (NA)	Wasserman 2004[43]	Hayes 2005[38] (1)	Hayes 2005[38] (2)	Kuo 2005[40] (3)	Kuo 2005[40] (1)	Kuo 2005[40] (2)	McReynolds 2007[41]	Cashel 1998[35]
Cost per QALY (two-stage strategy)	32,391	25,828	27,381	22,524	24,716	17,473	18,314	23,111	45,106

NA, not applicable.
a Here, the gold standard (e.g. DISC) is a single-stage screen and is included in the table to provide the reference standard.

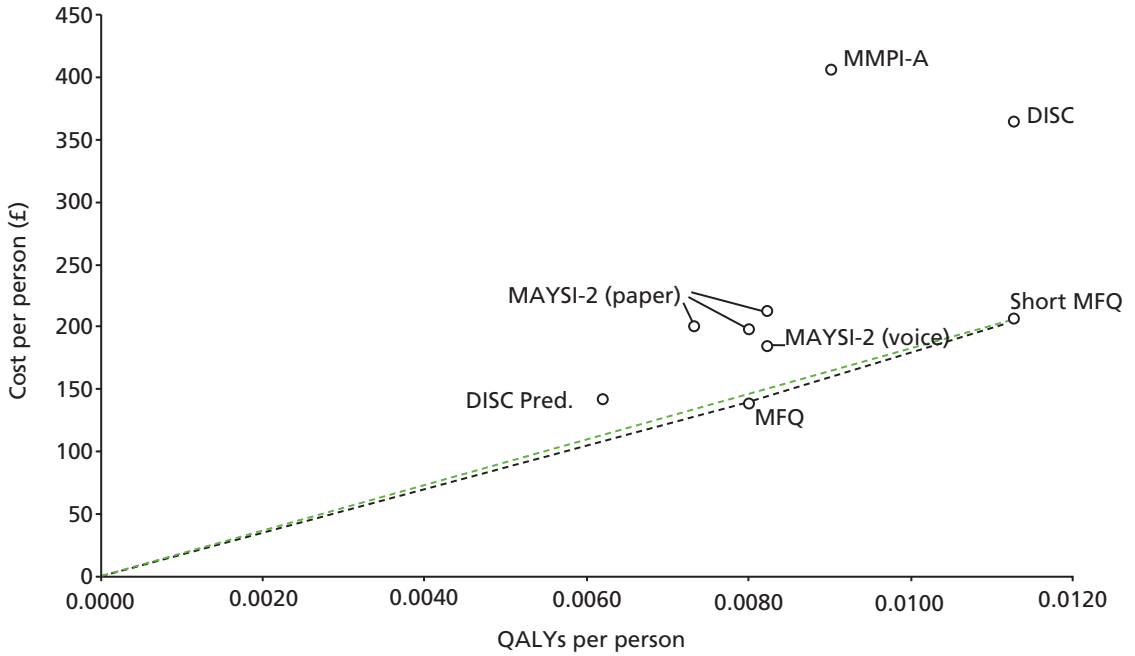


FIGURE 5 Cost-effectiveness plane of a two-stage screen detection strategy (single screen + gold standard) including the cost offset by reductions in recidivism attributed to treatment of depression. DISC Pred., the DISC predictive scales (DPS).

Although the exemplar approach provides a framework for future cost-effectiveness analyses of screening strategies in young offenders, the analyses presented so far have several limitations, in particular the limited data available with which to construct a model. The following section presents the results of sensitivity analyses to illustrate how sensitive the model is to variation in the input parameters.

Sensitivity analysis

Given the uncertainty in the current literature surrounding the real-world practice of the detection and treatment of mental health needs, this section presents deterministic sensitivity analysis of the input parameters driving the model. The impacts of variation in three parameters are presented, namely the prevalence of undetected depression under usual care; utility values and the severity of depression; and the level of effectiveness of CBT for depression on reducing recidivism.

Prevalence rate of undetected depression under usual care

The model utilises gender-specific prevalence rates of depression in young offenders in the UK as reported by Fazel *et al.*[2] The base-case exemplar uses these prevalence rates, which suggest that, on aggregate, 15% of all individuals in the system are depressed. However, the model assumes that, in the absence of an active detection strategy (single or two stage), all depressed individuals will go undetected and, therefore, untreated.

In the real-world setting, the prevalence rate under usual care will be made up of previously detected cases and currently undetected cases. *Table 25* illustrates how variation in the prevalence of depression driving the model alters the level of cost-effectiveness under the intersectoral perspective.

Table 25 suggests that, as the prevalence of depression in the young offending population increases (compared with the base-case prevalence of 15%), the ICER becomes smaller and vice versa. However, the analysis suggests that the change in the ICER is relatively small and the strategies that were cost-effective in the base-case analysis remain cost-effective in the sensitivity analysis, assuming a WTP threshold of £20,000–30,000 per QALY.

TABLE 25 Results of sensitivity analysis [cost per QALY (£)]: prevalence of depression (two-stage screening strategy, intersectoral perspective)

Model scenario (variation in prevalence rate)	Gold standard[a]	MAYSI-2				MFQ	Short MFQ	DPS	MMPI-A
		Paper	Paper	Voice	Paper				
Study (no. of screening tests in the study)	e.g. DISC (NA)	Wasserman 2004[43]	Hayes 2005[38] (1)	Hayes 2005[38] (2)	Kuo 2005[40] (3)	Kuo 2005[40] (1)	Kuo 2005[40] (2)	McReynolds 2007[41]	Cashel 1998[35]
Prevalence: 25%	30,077	24,263	25,658	21,311	23,272	16,778	17,507	21,887	41,865
Prevalence: 20%	31,178	25,008	26,478	21,889	23,959	17,109	17,891	22,469	43,408
Prevalence: 15% (base case)	32,391	25,828	27,381	22,524	24,716	17,473	18,314	23,111	45,106
Prevalence: 10%	33,732	26,735	28,380	23,228	25,554	17,877	18,783	23,821	46,984
Prevalence: 5%	35,224	27,744	29,491	24,011	26,486	18,327	19,304	24,610	49,072

NA, not applicable.
a Here, the gold standard (e.g. DISC) is a single-stage screen and is included in the table to provide the reference standard.

Utility values and the severity of depression

In applying the utility value to DFDs in the model to obtain QALYs, the base-case analysis is conservative in estimating the potential gains in identifying and treating depression because it assumes that all depressed individuals have mild depression. In reality, there is likely to exist a mixed picture of depression severity in the youth justice system; future research will benefit from better understanding levels of severity of depression in this specific population.

Tables 26–28 examine the effect of varying this assumption for all three scenarios previously discussed in the main results (single screen, two-stage screen and two-stage screen including the cost offset from reduced recidivism respectively).

Overall, these analyses suggest that the cost-effectiveness of all of the screening strategies across the three scenarios in *Tables 26–28* is highly sensitive to assumptions about the severity of depression. As expected, as the level of moderate depression increases (with a constant overall prevalence), all screening strategies become more cost-effective. For example, if the prevalence of mild to moderate depression is 40 : 60, the scenario for the use of single screening tools alone indicates that the use of the MFQ or the gold standard is likely to fall within the decision-maker's WTP threshold.

In part, these results are a reflection of the limitations of the available evidence on screening and treatment parameters. These results highlight the need for more research to measure quality of life and the severity of depression in young offenders, all of which are critical to address the current decision uncertainty.

Cognitive–behavioural therapy, rates of recidivism and consumption value of the quality-adjusted life-year

As the base-case results of the model suggest, potential additional gains through offsetting the cost of the discussed strategies may be highly influential in the cost-effectiveness analysis. However, two key assumptions have been made to present the base-case analysis, namely the consumption value of the QALY in adjusting cost offsets and the level of effect of CBT on reducing recidivism rates in the depressed population.

A decision-maker may apply the WTP threshold of £20,000–30,000 per QALY and a consumption value of health of £60,000 to down-weight the expected cost of crime averted. This therefore assumes that the cost of crime averted is divisible by three to estimate intersectoral cost-effectiveness.

The base-case exemplar assumes a 3 : 1 ratio for the consumption value of health. It is also feasible that, with greater certainty, a decision-maker's WTP may lie at the upper bound of £30,000 per QALY, implying that the cost offset should alternatively be down-weighted by a ratio of 2 : 1. *Table 29* provides a comparison of the effects of varying the assumed consumption value of health.

If a decision-maker working under greater certainty used the £30,000 per QALY threshold (implying that the non-health cost offset should be adjusted on a ratio of 2 : 1), a larger proportion of the two-stage detection strategies fall under the WTP threshold. This raises an interesting methodological question about weighting non-health costs and benefits when taking a broader intersectoral or societal perspective in evaluating health-care programmes (as in the case of the detection and treatment of mental health in young offenders).

The model uses the estimate of the effect of CBT on reducing recidivism as reported in the systematic review by Lipsey *et al.*[85] This may present an optimistic additional expectation from treatment and is higher than the expected benefit of treatment for depression as reported in Rohde *et al.*[64] and it may be more realistic to assume that the level of effect on reoffending may actually be lower.

TABLE 26 Results of sensitivity analysis [cost per QALY (£)]: ratio of mild to moderate depression – single screen (health-care perspective)

Model scenario (variation in severity of depression)	Gold standard	MAYSI-2				MFQ	Short MFQ	DPS	MMPI-A
		Paper	Paper	Voice	Paper				
Study (no. of screening tests in the study)	e.g. DISC (NA)	Wasserman 2004[43]	Hayes 2005[38] (1)	Hayes 2005[38] (2)	Kuo 2005[40] (3)	Kuo 2005[40] (1)	Kuo 2005[40] (2)	McReynolds 2007[41]	Cashel 1998[35]
Mild 100%, moderate (0%)	43,195	77,378	83,156	61,844	71,804	42,839	52,174	52,322	78,662
Mild 80%, moderate 20%	25,583	45,829	49,251	36,628	42,527	25,372	30,901	30,988	46,589
Mild 60%, moderate 40%	18,173	32,555	34,986	26,019	30,210	18,023	21,951	22,013	33,095
Mild 40%, moderate 60%	14,092	25,243	27,128	20,176	23,425	13,976	17,021	17,069	25,662
Mild 20%, moderate 80%	11,507	20,614	22,153	16,475	19,129	11,412	13,899	13,939	20,956
Mild 0%, moderate 100%	9724	17,419	18,720	13,922	16,164	9644	11,745	11,778	17,708

NA, not applicable.

TABLE 27 Results of sensitivity analysis [cost per QALY (£)]: ratio of mild to moderate depression – two-stage screening strategy (health-care perspective)

Model scenario (variation in prevalence)	Gold standard[a]	MAYSI-2				MFQ	Short MFQ	DPS	MMPI-A
		Paper	Paper	Voice	Paper				
Study (no. of screening tests in the study)	e.g. DISC (NA)	Wasserman 2004[43]	Hayes 2005[38] (1)	Hayes 2005[38] (2)	Kuo 2005[40] (3)	Kuo 2005[40] (1)	Kuo 2005[40] (2)	McReynolds 2007[41]	Cashel 1998[35]
Mild 100%, moderate (0%)	43,195	36,632	38,185	33,329	35,521	28,278	29,118	33,915	55,910
Mild 80%, moderate 20%	25,583	21,696	22,616	19,739	21,038	16,748	17,246	20,087	33,114
Mild 60%, moderate 40%	18,173	15,412	16,065	14,022	14,944	11,897	12,251	14,269	23,523
Mild 40%, moderate 60%	14,092	11,951	12,457	10,873	11,588	9225	9499	11,064	18,240
Mild 20%, moderate 80%	11,507	9759	10,173	8879	9463	7533	7757	9035	14,895
Mild 0%, moderate 100%	9724	8246	8596	7503	7996	6366	6555	7635	12,586

NA, not applicable.
a Here, the gold standard (e.g. DISC) is a single-stage screen and is included in the table to provide the reference standard.

TABLE 28 Results of the sensitivity analysis [cost per QALY (£)]: ratio of mild to moderate depression – two-stage screening strategy (intersectoral perspective)

Model scenario (variation in prevalence)	Gold standard[a]	MAYSI-2				MFQ	Short MFQ	DPS	MMPI-A
		Paper	Paper	Voice	Paper				
Study (no. of screening tests in the study)	e.g. DISC (NA)	Wasserman 2004[43]	Hayes 2005[38] (1)	Hayes 2005[38] (2)	Kuo 2005[40] (3)	Kuo 2005[40] (1)	Kuo 2005[40] (2)	McReynolds 2007[41]	Cashel 1998[35]
Mild 100%, moderate (0%)	32,391	25,828	27,381	22,524	24,716	17,473	18,314	23,111	45,106
Mild 80%, moderate 20%	19,184	15,297	16,217	13,340	14,639	10,349	10,847	13,688	26,715
Mild 60%, moderate 40%	13,628	10,866	11,520	9476	10,399	7352	7705	9723	18,977
Mild 40%, moderate 60%	10,567	8426	8933	7348	8063	5700	5975	7539	14,715
Mild 20%, moderate 80%	8629	6881	7294	6001	6585	4655	4879	6157	12,016
Mild 0%, moderate 100%	7292	5814	6164	5071	5564	3934	4123	5203	10,154

NA, not applicable.
a Here, the gold standard (e.g. DISC) is a single-stage screen and is included in the table to provide the reference standard.

TABLE 29 Results of sensitivity analysis [cost per QALY (£)]: consumption value of health – two-stage screening strategy (intersectoral perspective)

Model scenario	Gold standard[a]	MAYSI-2				MFQ	Short MFQ	DPS	MMPI-A
		Paper	Paper	Voice	Paper				
Study (no. of screening tests in the study)	e.g. DISC (NA)	Wasserman 2004[43]	Hayes 2005[38] (1)	Hayes 2005[38] (2)	Kuo 2005[40] (3)	Kuo 2005[40] (1)	Kuo 2005[40] (2)	McReynolds 2007[41]	Cashel 1998[35]
Consumption value of QALY 3 : 1 ratio (base case)	32,391	25,828	27,381	22,524	24,716	17,473	18,314	23,111	45,106
Consumption value of QALY 2 : 1 ratio	26,989	20,425	21,979	17,122	19,314	12,071	12,912	17,708	39,704

NA, not applicable.
a Here, the gold standard (e.g. DISC) is a single-stage screen and is included in the table to provide the reference standard.

The odds ratio for the treatment effect of CBT on recidivism in the base-case analysis was 1.53.[85] *Tables 30* and *31* present sensitivity analyses in which the odds ratio of the treatment effect of CBT for depression on recidivism was varied (for both the 3 : 1 and the 2 : 1 consumption values of health respectively).

Overall, this sensitivity analysis illustrates that the level of effect of CBT on recidivism is a major driver of whether or not strategies falls within the WTP threshold. As the odds ratio decreases (i.e. becomes closer to 1), the cost-effectiveness ratio increases. For instance, in the analysis with a consumption value of health of 3 : 1 (see *Table 30*), MFQ and Short MFQ in a two-stage strategy are likely to be cost-effective assuming the full effect of CBT on recidivism (1.53). However, small changes in the odds ratio alter the cost-effectiveness ratio so that it is now close to the upper bound of the WTP threshold. *Table 31* assumes that a decision-maker's WTP threshold is £30,000 per QALY and implies that a larger proportion of strategies may be cost-effective.

In the base-case scenario, the staff time for conducting the indicated screening programme within the criminal justice settings was costed using *Unit Costs in Criminal Justice* (i.e. 'Prisons: Nurse (mental health)'] at £24 per hour.[87] However, compared with *Unit Costs of Health and Social Care*[86] [i.e. Nurse (Mental Health) at a cost of £35 per hour], the base-case staff cost may be considered conservative. *Table 32* illustrates the effect of using the higher cost for staff time using a two-stage screening strategy and an intersectoral perspective.

Utilities for various depression states were required to inform the DFD approach to estimating QALYs for the treatment model. Revicki and Wood[81] provided the most suitable study having ascertained utilities using standard gamble interviews. Although these data are favourable in describing various states of depression, they have limitations in that the data are not UK or adolescent specific. Byford *et al.*[92] compared selective serotonin reuptake inhibitors and routine specialist care with and without CBT in adolescents with major depression. At baseline, this study provides a measure of utility (0.5) from a sample of 208 adolescents, aged 11–17 years, with moderate to severe major or probable major depression. *Table 33* illustrates how this utility would alter the base-case results.

Discussion

The cost-effectiveness analysis was conducted within the limitations of the available evidence on the effectiveness of screening and treatment strategies for mental health conditions in young offenders. Because of the limited evidence, we developed an exemplar model for depression to evaluate the cost-effectiveness of single-screening and two-stage screening followed by a treatment decision based on the screening outcome. Depression was chosen for the exemplar analysis for the reasons highlighted in the introduction to this chapter.

The limitations of the data are considerable and the results of the exemplar model must be interpreted within the context of these. This includes limitations in data availability for both screening and treatment parameters. For example, data on treatment were limited to a single study. In addition, the lack of data for a number of additional key input parameters, such as the impact of treatment on the rates of recidivism, limits the implications that can be drawn from the exemplar decision model.

The economic model identified key drivers required for decision analysis, which included the prevalence and severity of the mental health condition, the diagnostic accuracy of the screening instruments, the treatment effect on health outcomes and recidivism and the perspective of the economic analysis. The exemplar analysis demonstrated how strongly these key drivers could influence the cost-effectiveness decision. These key parameters are likely to be influential when further cost-effectiveness analyses are conducted in this population (whether for depression or other mental health conditions); hence, the cost-effectiveness analysis also informs areas of research prioritisation to reduce uncertainty in the current evidence base to allow evaluation of screening strategies in the future.

TABLE 30 Results of sensitivity analysis [cost per QALY (£)]: odds ratio of the treatment effect of CBT for depression on recidivism (assuming a ratio for the consumption value of health of 3 : 1) – two-stage screening strategy (intersectoral perspective)

		MAYSI-2							
Model scenario	Gold standard	Paper	Paper	Voice	Paper	MFQ	Short MFQ	DPS	MMPI-A
Study (no. of screening tests in the study)	e.g. DISC (NA)	Wasserman 2004[43]	Hayes 2005[38] (1)	Hayes 2005[38] (2)	Kuo 2005[40] (3)	Kuo 2005[40] (1)	Kuo 2005[40] (2)	McReynolds 2007[41]	Cashel 1998[35]
Odds ratio: 1.53 (base case)	32,391	25,828	27,381	22,524	24,716	17,473	18,314	23,111	45,106
Odds ratio: 1.4	34,557	27,993	29,547	24,690	26,882	19,639	20,480	25,276	47,272
Odds ratio: 1.3	36,406	29,843	31,396	26,540	28,732	21,489	22,329	27,126	49,121
Odds ratio: 1.2	38,442	31,878	33,431	28,575	30,767	23,524	24,364	29,161	51,157
Odds ratio: 1.1	40,693	34,129	35,682	30,826	33,018	25,775	26,615	31,412	53,408
Odds ratio: 1.0	43,195	36,632	38,185	33,329	35,521	28,278	29,118	33,915	55,910

NA, not applicable.

TABLE 31 Results of sensitivity analysis [cost per QALY (£)]: odds ratio of the treatment effect of CBT for depression on recidivism (assuming a ratio for the consumption value of health of 2 : 1) – two-stage screening strategy (intersectoral perspective)

		MAYSI-2							
Model scenario	Gold standard	Paper	Paper	Voice	Paper	MFQ	Short MFQ	DPS	MMPI-A
Study (no. of screening tests in the study)	e.g. DISC (NA)	Wasserman 2004[43]	Hayes 2005[38] (1)	Hayes 2005[38] (2)	Kuo 2005[40] (3)	Kuo 2005[40] (1)	Kuo 2005[40] (2)	McReynolds 2007[41]	Cashel 1998[35]
Odd ratio: 1.53 (base case)	26,989	20,425	21,979	17,122	19,314	12,071	12,912	17,708	39,704
Odd ratio: 1.4	30,238	23,674	25,227	20,371	22,563	15,320	16,160	20,957	42,953
Odd ratio: 1.3	33,012	26,448	28,001	23,145	25,337	18,094	18,935	23,731	45,727
Odd ratio: 1.2	36,065	29,502	31,055	26,198	28,390	21,147	21,988	26,784	48,780
Odd ratio: 1.1	39,441	32,878	34,431	29,575	31,767	24,524	25,364	30,161	52,156
Odd ratio: 1.0	43,195	36,632	38,185	33,329	35,521	28,278	29,118	33,915	55,910

NA, not applicable.

TABLE 32 Results of sensitivity analysis [cost per QALY (£)]: personnel cost for screening – two-stage screening strategy (intersectoral perspective)

		MAYSI-2							
Model scenario	Gold standard	Paper	Paper	Voice	Paper	MFQ	Short MFQ	DPS	MMPI-A
Study (no. of screening tests in the study)	e.g. DISC (NA)	Wasserman 2004[43]	Hayes 2005[38] (1)	Hayes 2005[38] (2)	Kuo 2005[40] (3)	Kuo 2005[40] (1)	Kuo 2005[40] (2)	McReynolds 2007[41]	Cashel 1998[35]
Staff cost £24 per hour (base case)	32,391	25,828	27,381	22,524	24,716	17,473	18,314	23,111	45,106
Staff cost £35 per hour	44,149	34,577	36,842	29,760	32,957	22,394	23,619	30,615	62,691

NA, not applicable.

TABLE 33 Results of the sensitivity analysis [cost per QALY (£)]: variation in utility referencing – two-stage screening strategy (intersectoral perspective)

		MAYSI-2							
Model scenario	Gold standard	Paper	Paper	Voice	Paper	MFQ	Short MFQ	DPS	MMPI-A
Study (no. of screening tests in the study)	e.g. DISC (NA)	Wasserman 2004[43]	Hayes 2005[38] (1)	Hayes 2005[38] (2)	Kuo 2005[40] (3)	Kuo 2005[40] (1)	Kuo 2005[40] (2)	McReynolds 2007[41]	Cashel 1998[35]
Utility for depressed days (mild severity): 0.64[81] (base case)	32,391	25,828	27,381	22,524	24,716	17,473	18,314	23,111	45,106
Utility for depressed days (mild severity): 0.50[92]	22,743	18,134	19,225	15,815	17,354	12,269	12,859	16,227	31,670

NA, not applicable.

The economic analysis also highlighted the value of having an effective treatment for any screening strategy to be cost-effective. For the exemplar model, our systematic review found only one relevant study (i.e. Rohde *et al.*[64]) that could be used to derive DFDs and in turn QALYs. The treatment model was based on data from this single study, with a relatively small sample size of only 93 adolescents; moreover, although the recovery functions used to derive DFDs showed small gains after CBT, these functions were not statistically significantly different from each other. Moreover, using point estimates of recovery rates from the Rohde *et al.* study,[64] the treatment model showed that, if all depressed young offenders could be treated with CBT (assuming that they could be identified at no cost), the cost per QALY was £17,542. Given that screening strategies are imperfect and result in unnecessary treatment costs because of false-positive individuals and missing out potential health gains for false-negative individuals, the cost per QALY estimates would only become higher, in turn making screening less cost-effective. This puts the emphasis on finding effective evidence-based treatment strategies that would produce reasonable health benefits after identifying screen-positive individuals. This could be achieved by conducting larger, high-quality clinical trials to identify or develop effective treatment strategies for screen-positive individuals.

The cost-effectiveness analysis also raised interesting methodological issues around the perspective of the economic analysis and the valuing of non-health benefits and costs for health services decision-making. The exemplar model suggests that several screening strategies that were not likely to be cost-effective from the health services perspective (given the available evidence) may become cost-effective when an intersectoral perspective is adopted. Moreover, the consumption value of a QALY used in the analysis was found to be another important determinant in the cost-effectiveness analysis.

The evidence on the sensitivity and specificity of screening instruments used in the exemplar model shows that either most instruments have poor diagnostic ability or there is significant uncertainty around sensitivity and specificity or both. Our analysis suggested that, if an instrument produces a high number of false-positive results, introducing two-stage screening (with the second stage being the gold standard) may be more cost-effective if the second screening cost could offset the cost of incorrectly treating the false positives. Although uncertainty in the available evidence does not allow us to reach definite conclusions, the exemplar model indicates that in the presence of poor diagnostic properties a two-stage screening strategy may be more cost-effective than single-stage screening.

In conclusion, the cost-effectiveness analysis presented here is primarily an exemplar. It provides an insight into the decision problem and identifies key drivers of cost-effectiveness and demonstrates the effects of our level of uncertainty about model parameters. This is of use in informing future research priorities, which will be discussed in *Chapter 10*. Before these future research priorities are identified, the current evidence base is assessed against UK NSC criteria.

Chapter 9 Evaluation of the current evidence base against UK National Screening Committee criteria

The UK NSC was founded in 1996 with the remit of providing advice to ministers and the NHS about the value of screening for a range of health conditions. The aim of the committee is to ensure that screening in the UK does more good than harm and that quality is ensured at each step of a screening programme. In providing recommendations, the UK NSC draws on a wide range of evidence to evaluate the likely benefits of a particular screening programme and does so through an assessment of the evidence for screening against a number of internationally recognised criteria.

There are currently no UK NSC recommendations about the value of screening for mental health problems in young people who offend; in fact, current recommendations for mental health problems in the wider population are limited.

The aim of this chapter was to answer the following research question: 'Do current screening strategies for mental health problems in young people who offend meet minimum criteria laid down by the UK NSC'?

The UK NSC criteria for screening are summarised in *Box 1*. The results of the evidence syntheses conducted as part of this review are relevant to five of the criteria (5, 6, 10, 13 and 16).

BOX 1 Summary of UK NSC criteria

The condition

1. The condition should be an important health problem.
2. The epidemiology and natural history of the condition, including development from latent to declared disease, should be adequately understood and there should be a detectable risk factor, disease marker, latent period or early symptomatic stage.
3. All of the cost-effective primary prevention interventions should have been implemented as far as practicable.
4. If the carriers of a mutation are identified as a result of screening, the natural history of people with this status should be understood, including the psychological implications.

The test

5. There should be a simple, safe, precise and validated screening test.
6. The distribution of test values in the target population should be known and a suitable cut-off level defined and agreed.
7. The test should be acceptable to the population.
8. There should be an agreed policy on the further diagnostic investigation of individuals with a positive test result and on the choices available to those individuals.
9. If the test is for mutations the criteria used to select the subset of mutations to be covered by screening, if all possible mutations are not being tested, should be clearly set out.

BOX 1 Summary of UK NSC criteria (*continued*)

The treatment

10. There should be an effective treatment or intervention for patients identified through early detection, with evidence of early treatment leading to better outcomes than late treatment.
11. There should be agreed evidence-based policies covering which individuals should be offered treatment and the appropriate treatment to be offered.
12. Clinical management of the condition and patient outcomes should be optimised in all health-care providers before participation in a screening programme.

The screening programme

13. There should be evidence from high-quality randomised controlled trials that the screening programme is effective in reducing mortality or morbidity.
14. There should be evidence that the complete screening programme (test, diagnostic procedures, treatment/intervention) is clinically, socially and ethically acceptable to health professionals and the public.
15. The benefit from the screening programme should outweigh the physical and psychological harm (caused by the test, diagnostic procedures and treatment).
16. The opportunity cost of the screening programme (including testing, diagnosis and treatment, administration, training and quality assurance) should be economically balanced in relation to expenditure on medical care as a whole (i.e. value for money). Assessment against this criterion should have regard to evidence from cost–benefit and/or cost-effectiveness analyses and have regard to the effective use of available resource.
17. All other options for managing the condition should have been considered (e.g. improving treatment, providing other services) to ensure that no more cost-effective intervention could be introduced or current interventions increased within the resources available.
18. There should be a plan for managing and monitoring the screening programme and an agreed set of quality assurance standards.
19. Adequate staffing and facilities for testing, diagnosis, treatment and programme management should be available prior to the commencement of the screening programme.
20. Evidence-based information, explaining the consequences of testing, investigation and treatment, should be made available to potential participants to assist them in making an informed choice.
21. Public pressure for widening the eligibility criteria, for reducing the screening interval and for increasing the sensitivity of the testing process should be anticipated. Decisions about these parameters should be scientifically justifiable to the public.
22. If screening is for a mutation the programme should be acceptable to people identified as carriers and to other family members.

Criterion 5

'There should be a simple, safe, precise and validated screening tool.'

Chapter 4 examined the diagnostic test accuracy of existing screening tools, which relates to the 'precision' component of this criterion. In terms of the 'validation' component, *Chapter 4* also examined the extent to which validity data were available for those mental health needs assessments that did not report diagnostic test accuracy information.

Although a number of screening measures were examined for a range of mental health problems, there were too few studies of any one screening measure for any one mental health problem to firmly establish the precision of a measure in this population. The MAYSI-2[10] was the most widely examined of all of the

screening instruments but even for this measure there was an insufficient number of studies to conduct a diagnostic meta-analysis. Furthermore, the data for the MAYSI-2[10] suggested that at the recommended cut-off points the measure had only moderate sensitivity and specificity across a number of mental health domains.

We also found limited validation data for those measures for which diagnostic test accuracy information was not reported. Even were substantial validation information to be reported for these measures, they would still fail to meet UK NSC criterion 5 because it assumes that a screen is both valid and precise. Mental health needs assessments that do not provide information on diagnostic test accuracy would necessarily fail this criterion, because information on precision is not available.

Criterion met? No.

Criterion 6

> *'The distribution of test values in the target population should be known and a suitable cut-off level defined and agreed.'*

Studies typically reported the diagnostic performance of a particular screening measure at a limited range of cut-off points, so the distribution of test values in the target population is not known. However, studies did frequently report sensitivity and specificity at recommended or standard cut-off points commonly cited in the literature. For use as a screening measure, high sensitivity is typically protected at the expense of somewhat lowered specificity, although, as described in *Chapter 2*, a balance often needs to be struck between high sensitivity and moderate specificity. Often the sensitivity at these recommended cut-off points was lower than that which is typically required for use as a screening instrument. This did appear to be the case for the MAYSI-2,[10] for example. However, simply altering the cut-off point to increase sensitivity may not be possible given that, as described above, even at moderate levels of sensitivity the observed specificity was also moderate. Altering the cut-off point to increase sensitivity would inevitably further reduce specificity and this may lead to an unacceptably high false-positive rate. For the MAYSI-2 this may make it difficult to establish an agreed cut-off level, although any conclusions about the MAYSI-2 are limited by the small number of studies examining its performance for the same diagnostic category. For other instruments, for which the evidence base is even more limited, this conclusion also holds.

Criterion met? No.

Criterion 10

> *'There should be an effective treatment or intervention for patients identified through early detection, with evidence of early treatment leading to better outcomes than late treatment.'*

In *Chapter 6* a small number of studies were identified that examined the clinical effectiveness of psychological treatments for a range of mental health difficulties in young people who offend. No studies were identified that examined the effectiveness of psychopharmacological interventions.

Although some studies did report some positive findings about the potential benefits of psychological interventions in this population, caution is needed in interpreting these results. First, there were too few studies of any one intervention for any one mental health difficulty to make firm conclusions about the likely effectiveness of these interventions. Second, the quality assessment of the studies suggested that many had either a high risk of bias or an unclear risk of bias for a number of key methodological features. For example, allocation concealment, the absence of which is empirically associated with increased effect

sizes,[69] was rated as unclear or at high risk of bias in all but one of the 10 studies reviewed. Blinding of outcome assessment, the absence of which is also likely to be associated with inflated effect sizes, was rated as unclear or at high risk of bias in all but two studies.

The existing studies examine whether or not the interventions were effective in general; they do not specifically examine whether or not early treatment led to improved outcomes relative to later treatment.

On the basis of the current evidence base, therefore, it would be premature to conclude that there are effective interventions for mental health problems in this population in general. It is also unknown if early intervention is more effective than later treatment.

Criterion met? No.

Criterion 13

> *'There should be evidence from high-quality randomised controlled trials that the screening programme is effective in reducing mortality or morbidity.'*

We were unable to identify any studies that had examined the effectiveness of a mental health screening programme in a population of young people who offend.

Criterion met? No.

Criterion 16

> *'The opportunity cost of the screening programme (including testing, diagnosis and treatment, administration, training and quality assurance) should be economically balanced in relation to expenditure on medical care as a whole (i.e. value for money).'*

Chapter 8 produced an illustrative decision model to evaluate the opportunity costs of screening programmes for the identification of depression in young people who offend. This model provides a preliminary framework for evaluating the value for money of various screening strategies. However, there exists considerable uncertainty in the input parameters, which, as described in the next chapter, will require substantial future research before conclusions related to this criterion can be drawn.

Criterion met? No.

Summary

The UK NSC provides a detailed list of the criteria that should ideally be met before a screening programme is used in the UK. We were able to evaluate a number of these criteria on the basis of the results of the evidence syntheses. Of the criteria we were able to examine, none was currently met for any screening method for any of the mental health problems we examined as part of this review.

Chapter 10 Identifying priorities for future research

In *Chapter 2* the decision problem faced by the health and youth justice systems in the identification and treatment of mental health problems in young people who offend was identified. This chapter revisits the uncertainties around this decision problem. It begins by summarising the current state of the evidence, paying particular attention to remaining evidence gaps. In *Chapter 8* we developed a decision model for depression screening but recognise that this is based on extremely limited information and is best seen as an exemplar of the type of modelling that could be carried out in this area were a larger number of methodologically robust studies to be available. Given the limited nature of the evidence and the very provisional status of the model, there was insufficient information to carry out a formal value of information (VoI) analysis. However, the decision model can provide some detail on key drivers that are likely to be relevant to reducing decision uncertainty through future research. In the light of the evidence gaps identified by this review and the key drivers that emerge from the model, the chapter concludes by outlining key research priorities.

Evidence gaps identified by the systematic reviews

Table 34 summarises the current evidence base on screening for a range of mental health problems in young people who offend. The numbers refer to the number of studies identified by the review for each of the broad areas for which we sought to identify relevant evidence. When more than one study reported data from the same or an overlapping data set, this is counted as a single study in the table.

TABLE 34 Summary of the numbers of studies providing evidence relevant to the review

	Screening			Treatment	
Mental health difficulty	Diagnostic test accuracy	Clinical effectiveness of screening	Cost-effectiveness of screening	Clinical effectiveness	Cost-effectiveness
Mood disorders					
Major depression	2	0	0	1	0
Other/any depressive disorder	2	0	0	0	0
Anxiety disorders					
Generalised anxiety disorder	1	0	0	0	0
PTSD	1	0	0	1	0
Other/any anxiety disorder	2	0	0	0	0
Disruptive disorders					
ADHD	2	0	0	0	0
Conduct disorder	1	0	0	2	0
ODD	1	0	0	0	0
Any/other disruptive disorder	2	0	0	0	0
Other	0	0	0	0	0
General mental health problems	0	0	0	6	0

As *Table 32* indicates, there are more gaps than evidence, and when evidence does exist it is limited. For many of the mental health problems examined as part of this review, the evidence was restricted to one or two studies of diagnostic test accuracy; for a small number of problems there was additional information on the effectiveness of clinical interventions. For areas in which some information was available, quality assessments indicated that many of the studies were either at high or unclear risk of bias across a number of domains. The existing evidence therefore does not provide a robust evaluation of the test accuracy, clinical effectiveness or cost-effectiveness of screening or more generally the clinical effectiveness or cost-effectiveness of treatments for mental health problems among young people who offend.

Clearly, many uncertainties remain. Given the limited nature of the current evidence base, both in terms of quantity and quality, a general recommendation is to conduct an extensive programme of more fundamental work before a randomised controlled trial of screening for mental health problems in young people is considered. This more fundamental work should include methodologically robust trials of interventions for mental health problems in this population as well as comprehensive studies of diagnostic test accuracy. The number of relevant studies identified was small for both community and incarcerated settings so there is a need to conduct this more fundamental work in both of these settings.

Before providing detailed research recommendations, however, the insights gained from the decision model are outlined, because these help to more clearly specify the nature of future research priorities.

Insights from the decision model

Value of information analysis evaluates the opportunity cost arising from making a suboptimal (or wrong) decision (such as opting for a suboptimal screening and treatment strategy in this context) on the basis of imperfect current evidence. The underlying idea is that uncertainty in parameters results in uncertainty in a decision, which has an opportunity cost that can be reduced by carrying out further research. By subscribing a monetary value to reducing uncertainty in parameters, the VoI analysis indicates how much decision-makers should be willing to pay to optimise their decision. As conducting research can be expensive, its cost should be contrasted against the consequences of making a wrong decision. Hence, the VoI analysis informs whether or not future research is worthwhile (i.e. the potential 'payback' of expenditure on research) The VoI approach ensures that allocation of funds to health research is in proportion to the opportunity cost to the health services.

The expected value of perfect information (EVPI) estimates the expected total losses given decision uncertainty. Should the EVPI suggest that further research is warranted, a decision model can be utilised to indicate specific input parameters that contribute most to decision uncertainty (by parameter we mean the evidence used in the analysis presented in *Chapter 8*, such as the prevalence of mental health problems). The expected value of partial perfect information (EVPPI) can identify which specific parameter or set of parameters is likely to represent the best value for investment in research.

This review would have ideally liked to quantify the value of further research on mental health screening and treatment in young offenders. However, as we have discussed in previous chapters, the current evidence base to inform decision-making in the context of mental health in young offenders is very limited; therefore, it does not allow us to develop a comprehensive probabilistic framework to formally evaluate the opportunity cost of making a suboptimal decision because of uncertainty in parameters. Despite these limitations, an exemplar model for depression was developed using the limited available evidence, which allowed us to identify the key drivers of decision uncertainty. These key drivers and their potential impact were evaluated in a number of sensitivity analyses in *Chapter 8*. In the absence of formal VoI analysis, examination of these key drivers can help identify key priorities for future research. The exemplar model identified three key parameters that should be prioritised in future research, namely (1) the prevalence of unidentified mental health problems that in the absence of screening would go undetected; (2) the effectiveness of interventions to improve mental health in young offenders that can be evaluated using

generic measures (e.g. QALYs); and (3) the impact of interventions on recidivism. The remainder of this section explores each of these parameters individually by examining the assumptions made, evaluating the sensitivity of the decision to varying levels of input parameters and evaluating the potential impact of resolving uncertainty in these parameters in terms of the allocation decision.

The prevalence of unidentified mental health problems in usual care

The prevalence parameter is crucial in evaluating the incremental costs and benefits of screening and treatment compared with usual care. In the absence of evidence to establish patient pathways for undiagnosed mental health needs, the current exemplar model assumes that, without an active detection strategy, individuals will default into a group in which they will be undetected and therefore will not be treated accordingly. However, uncertainty in this parameter is an important limitation in conducting an incremental analysis against usual care within this population.

The potential impact of uncertainty in this parameter was evaluated in the sensitivity analysis by varying the prevalence of undetected depression in the exemplar model. The intention was to imply varying proportions of individuals who may be identified under usual care and thereby alter the proportion of new cases that could feasibly be detected through an active detection strategy.

The results suggest that the higher the prevalence of undetected depression, the more cost-effective the screening strategies would become. This implies that the value associated with the current uncertainty surrounding the prevalence of unidentified mental health problems in usual care is substantial and that there is value in future research reducing this uncertainty.

Effectiveness of interventions and uncertainty in benefits in terms of utility-based measures (e.g. quality-adjusted life-years)

The absence of any previous studies of cost-effectiveness in this area was discussed in *Chapter 7*. However, the exemplar model was able to use data from Rohde *et al.*[64] to map disease-specific outcomes onto a generic measure using DFDs and thereby estimate health outcomes in terms of QALYs. However, the Rohde *et al.* study[64] included a cohort with baseline disease severity measured as 'mild depression'.[81] The criminal justice system is likely to include adolescents with different levels of depression severity. The Rohde *et al.* study[64] does not allow us to estimate the treatment effect of CBT in moderate or severe depression. Moreover, the evidence in the Rohde *et al.* study[64] was based on a small sample size of 93 adolescents, with substantial uncertainty in the effectiveness of the intervention.

The baseline disease severity of the cohort may alter the cost-effectiveness of an identification strategy. In the base-case exemplar model it was assumed that the cohort included only mildly depressed individuals (in line with the Rohde *et al.* study[64]). Altering the underlying assumption to one in five individuals being moderately depressed means that all active detection strategies become more cost-effective (assuming a constant treatment effect). However, besides uncertainty in the underlying prevalence of depression, there is also uncertainty in how the severity of depression would interact with the treatment effect. Finally, as the Rohde *et al.* study[64] does not report outcomes in terms of utility measures, using an indirect approach to estimate QALYs further adds to the uncertainty in the treatment effect. Hence, including generic measures to estimate treatment benefit is likely to reduce decision uncertainty.

Impact of recidivism on the cost per quality-adjusted life-year

The exemplar model for the detection and treatment of depression in young offenders further considered intersectoral costs and benefits through estimating changes in recidivism. The model suggests that the extent to which treatment for mental health problems alters recidivism is highly influential for decision-making. Varying the odds ratio of the impact of treating depression on recidivism shows that, for every 10% decrease in the odds ratio, the cost-effectiveness ratio increases by 5–10%. However, there is lack of evidence on the relationship between the severity of depression and the recidivism-related treatment effect of CBT. As such, there is value in future trials of mental health treatments in this area to

collect specific information on subsequent criminal activity to establish the relationship between treatment for depression and recidivism.

Second, the economic analysis has highlighted that the costs and benefits of interventions with an intersectoral impact (such as on the health system or the youth justice system) may be valued differently by different decision-makers. Our analysis suggested that the weight associated with costs and benefits incurred outside of the health system is an important driver of decision uncertainty. Hence, further research on the consumption value of a QALY when evaluated from a societal perspective will have a significant impact on decision-making.

Summary

In summary, the sensitivity analyses surrounding key drivers in the exemplar model highlighted the areas that future research should prioritise to reduce decision uncertainty. The prevalence of untreated mental health problems in young offenders, evaluation of treatment effects in terms of utility-based measures and the effect of mental health interventions on the rate of recidivism are all found to be key drivers of the decision problem and should be incorporated into the future research agenda.

Description of future research priorities

The combination of the results of the systematic reviews, particularly the large evidence gaps identified, and the insights about key drivers offered by the decision model suggest a number of research priorities for diagnostic test accuracy studies, clinical effectiveness and cost-effectiveness trials of interventions and clinical effectiveness and cost-effectiveness trials of screening.

Recommendations for clinical effectiveness and cost-effectiveness trials of interventions

The first recommendation relates to future trials of the clinical effectiveness and cost-effectiveness of interventions for mental health problems in young people who offend, because screening is justifiable only if there exist effective treatments, something that is not yet established in this population.

The review identified a small number of predominantly small trials examining the clinical effectiveness of psychological interventions. Although it would be possible to consider these as feasibility trials, providing, for example, estimates of likely effect sizes, it is arguable that, with the exception of a small number of studies, the existing research has a number of methodological limitations and these may make it necessary to conduct further feasibility trials ahead of any definitive trials. Furthermore, there is limited evidence from a UK context and it may be necessary to establish the feasibility of conducting a trial in this setting. In fact, there are likely to be distinct challenges of conducting such trials in UK community and residential settings. Research in each is likely to be necessary given the current limited evidence base.

One of the roles of a feasibility trial is to help to clarify a number of important parameters ahead of large-scale definitive trials. Future feasibility trials in this area should seek to provide clarity about a number of parameters. This includes providing information on variance of the outcome measure to inform future power calculations. The risk of bias in many of the reviewed studies may have served to artificially inflate the observed effect of an intervention. Future work is therefore required to better establish the likely size of the effect of interventions in this area. It will also be necessary to establish attrition rates from the interventions, something that was often poorly reported in the studies evaluated as part of this review. In connection with this, it may be appropriate to include a qualitative component in these trials examining the acceptability of interventions to both the young people and the professionals involved in their care, both within the youth justice system and within health services. The cost-effectiveness of interventions may emerge over time and so it may be necessary to establish what length of follow-up is feasible in trials within this particular population.

The insights from the decision model about key drivers of uncertainty suggest a number of further recommendations. These include ensuring that future trials gather information to permit the calculation of QALYs. The model also suggests that the impact of the interventions for mental health problems on recidivism may be important in determining their cost-effectiveness. Future trials should therefore seek to establish the extent to which treatment alters intersectoral outcomes, including recidivism.

Recommendations for diagnostic test accuracy studies

There is a need for methodologically robust diagnostic test accuracy studies that validate available screening measures against a gold standard diagnostic interview conducted to internationally recognised criteria (e.g. DSM, ICD).

The decision problem considered a number of potential screening pathways but identified substantial uncertainties around the likely cost-effectiveness of these different pathways. This includes uncertainty about the accuracy of different classes of screening measures (e.g. bespoke measures for use in young offender populations vs. measures designed for use outside of this population). Within different classes there is also no clear indication that particular measures have superior operating characteristic relative to other measures. Future studies should therefore directly compare a number of available instruments across and within these broad categories. Current UK policy recommends the use of the CHAT mental health screen. Given that this is currently recommended, it should be incorporated into such evaluations, with its performance directly compared with the performance of a range of other measures from these broad classes of screening measures. From a UK perspective, it will be necessary to establish the diagnostic performance of the CHAT in both community and residential settings. The characteristics of young people in these two settings are likely to differ in a number of ways, which may affect the diagnostic performance of the tests. For example, it would be inappropriate to assume that the observed sensitivity and specificity in one setting will hold for another setting.

As the decision model emphasises, the cost-effectiveness of screening is influenced by the assumptions made about the prevalence of previously unidentified mental health problems. Future diagnostic test accuracy studies should seek to separately report the diagnostic performance of the measures for previously unidentified cases and all cases (combining previously identified and previously unidentified cases) of mental health problems in young people who offend. The studies should be adequately powered to ensure suitably narrow CIs around the sensitivity estimates given the likely prevalence of previously undetected mental health problems.

Recommendations for clinical effectiveness and cost-effectiveness trials of screening

The review identified no studies of examining the clinical effectiveness or cost-effectiveness of screening. In the absence of any previous research in this area, and in light of the substantial limitations of more fundamental work on diagnostic test accuracy and clinical effectiveness, it would be premature to consider a trial of screening for mental health problem in young people who offend. Instead, it may be more appropriate to conduct further decision modelling once this more fundamental work has been conducted to better inform the nature of and need for future screening trials.

Summary

The series of systematic reviews and insights from the decision model suggest a number of research priorities. These include the need for methodologically robust trials of interventions that permit the calculation of QALYs and assess the impact of interventions on recidivism. These also include the need for methodologically robust diagnostic test accuracy studies that provide an indication of the accuracy of screening instruments in identifying previously unidentified mental health problems. It may be appropriate to answer these more fundamental research questions before trials of screening are undertaken. Research in both community and incarcerated settings is equally limited and so the same research priorities apply equally to both.

Chapter 11 Discussion

Mental health problems are common among young people who offend and are linked to a range of negative consequences, including increased rates of recidivism. These problems, however, remain under-recognised and undertreated. In recognition of this, policy strategies have recommended the use of screening for such problems. The aim of this review was to establish the value of screening in this population and the groups in whom this may be of most benefit. This aim was divided into a number of objectives; the results are briefly summarised for each of these.

Statement of principal findings

Objective 1: to conduct a systematic review and evidence synthesis of the diagnostic properties and validity of existing screening measures for mental health problems in young people who offend

The review identified nine relevant studies, eight of which examined diagnostic test accuracy and one of which examined the validity of mental health screening methods.

There was an insufficient number of studies to make firm conclusions about the accuracy of screening measures in young people who offend and quality assessment also indicated a high or unclear risk of bias for many studies across a number of domains. Any conclusions are therefore necessarily tentative. The MAYSI-2[10] was the most widely evaluated of the measures although, even here, the number of studies was small – too small, for example, to use diagnostic meta-analytic techniques. At literature standard cut-off points the MAYSI-2[10] typically had moderate sensitivity and specificity; its value, therefore, as a screening measure is not clear. There was also no evidence that the screening accuracy of this measure, a measure specifically designed for use in groups of young people who offend, was superior to that of other general measures.

For those screening instruments described as mental health needs assessments for which we were unable to identify studies relating to diagnostic test accuracy, we sought evidence relating to the validity of the assessments. Even were evidence to be identified, it is not necessarily clear how this would be integrated with standard health services research methods of evaluating the cost-effectiveness of screening. As it was, we identified very limited evidence on the validity of these screening measures.

Objective 2: to assess the clinical effectiveness of screening strategies in this population and (more broadly) to assess the clinical effectiveness of interventions for mental health problems

No studies were identified that examined the clinical effectiveness of screening for mental health problems in young people who offend, but 10 randomised controlled trials were identified that examined the clinical effectiveness of interventions for mental health problems. Studies examined interventions for depression, anxiety, including PTSD, conduct disorder, ODD and ADHD. The majority of studies examined had a broader focus, such as improving interpersonal functioning. CBT was the most commonly evaluated intervention.

There were too few studies to make firm conclusions about the clinical effectiveness of any single intervention for any single mental health problem, particularly because the quality assessment suggested that many of the studies were rated as being at high or unclear risk of bias across a number of domains. These included biases such as allocation concealment and blinding of outcome assessment, the presence of which are likely to artificially inflate observed effect sizes. Two studies, however, were rated as being at low risk of bias across a number of domains.[56,64] The clinical effectiveness of interventions for the mental health problems examined as part of this review is currently unknown.

Objective 3: to assess the cost-effectiveness of screening strategies in this population and (more broadly) to assess the cost-effectiveness of interventions for mental health problems

The review identified no studies of the cost-effectiveness of screening or interventions for mental health problems.

To evaluate the cost-effectiveness of identification strategies, the policy question addressed by the decision model was constrained to focus on the screening and subsequent management of one common mental health problem in the young offender population: depression. The decision model provides initial insights into the possible merits of identification and treatment strategies and the importance of perspectives adopted given the intersectoral nature of this question. However, these insights need to be considered within the limitations of the available evidence emerging from the systematic review of diagnostic and clinical effectiveness studies. Nonetheless, the decision model makes a contribution to the overall evidence by providing an exemplar based on a formal quantitative framework that provides an indication of the various inputs and data sources required to appropriately inform a cost-effectiveness assessment.

Objective 4: to assess whether or not current screening strategies meet minimum criteria laid down by the UK National Screening Committee

The earlier phases of the systematic review provided evidence relevant to five of the UK NSC criteria. These included the existence of a precise and valid screening instrument (criterion 5), a known distribution of test values and an agreed cut-off for the instrument (criterion 6), the existence of an effective treatment (criterion 10), evidence from randomised controlled trials that screening is effective (criterion 13) and opportunity costs should be economically balanced in relation to expenditure on medical care (criterion 16). On the basis of the existing evidence we concluded that none of the five criteria was currently met.

Objective 5: to identify research priorities and the value of developing future research into screening strategies for young offenders with mental health problems

There was insufficient evidence from the earlier phases of the review to formally conduct a VoI analysis. However, on the basis of the identified evidence gaps and insights from the exemplar decision model a number of research priorities were identified, which are summarised in *Chapter 12*.

Limitations

The results of the current review should be interpreted in the light of limitations of the review itself and limitations of the primary studies. Although the next chapter will summarise the main suggested research priorities, this section on limitations will also offer recommendations for improving the methodological quality and the quality of reporting of future studies in this area.

Limitations of the current review

The search strategy we developed for the review used terms to identify young people. The ideal strategy would be to not limit the search by age group, because indexing and the use of age-related terms in the titles and abstracts of database records is often poor. However, it was agreed that the number of records retrieved from unlimited age group searches was unmanageable and that the concept of 'young offenders' is well understood and recognised in the criminal justice and forensic field and so would be more likely to be included in the titles, abstracts and indexing of database records.

The search for validation studies of mental health needs assessments may have missed important studies because the identification strategy relied on a reference to the measure in the title or abstract along with a reference to validity data. As described in more detail in *Chapter 4*, it is possible that relevant validity data may be contained in papers that did not reference the measure in the abstract in this way.

Alcohol and drug problems have a higher prevalence in young people who offend than in the general population, but this was not examined as part of the review. Our review cannot, therefore, address the extent to which screening for drug and alcohol problems may be clinically effective or cost-effective, or the effect of screening for these problems alongside screening for mental health problems.

We did not formally seek to review the evidence on the acceptability to young people who offend of either the screening methods or the interventions. Establishing the acceptability of a screening strategy in the population in which it is intended to be used is important because it may have a substantial impact on the effectiveness of that strategy. Acceptability is in fact one of the criteria used by the UK NSC in deciding whether or not a screening programme should be recommended for use in the UK. Feedback from the patient and public involvement group did indicate some preferences for shorter screening measures and talking treatments rather than medication, although further work in this area is needed.

The searches were carried out in April 2011 and it is possible that subsequent publications would alter the conclusions of the review. In particular, we are aware that diagnostic test accuracy data for the CHAT, including the mental health section, have recently been published.[23] Future reviews of diagnostic test accuracy and future decision modelling should seek to incorporate these data.

The approach of creating an exemplar decision model was taken because of the very limited amount of literature currently available in this specific area to inform the decision-maker. As such, there are significant limitations in the parameter inputs that need to be considered. First, the treatment model is based on a single study with a small sample size;[64] considerable caution is therefore needed in making any inferences about QALYs. Second, the input parameters for diagnosis are point estimates and have not taken into account the CIs to reflect the level of uncertainty. Finally, the lack of certainty surrounding key input parameters (i.e. the prevalence of unidentified mental health problems, measures of utility, the impact of mental health treatment on recidivism) limits the potential conclusions that can be drawn from this exemplar case study.

Limitations of the primary studies

Limitations of the diagnostic test accuracy studies

The QUADAS-2 tool[33] was used to evaluate the quality of the diagnostic test accuracy primary studies. The risk of bias of the included diagnostic test accuracy studies was rated as unclear or high for many of the studies across most of the bias domains with the exception of the reference standard domain.

The frequency with which an unclear rating was given suggests that future studies in this area should more clearly report key methodological features that are likely sources of bias. Future studies should ensure that sufficient information is reported to enable each item on the QUADAS-2 tool to be assessed. This includes, for example, clear statements about whether or not the index test and the reference standard were interpreted blind to each other, the length of time between the administration of the index test and the administration of the reference test and the flow of participants through the study. The Standards for the Reporting of Diagnostic Accuracy Studies (STARD) statement[93] also provides guidelines for the reporting of diagnostic test accuracy studies and these should also be considered when reporting test accuracy in this area.

In particular, future studies should report the performance of the screening measures at all cut-off points to prevent the post hoc selection of cut-off points and so that future modelling can examine the effect of different balances between sensitivity and specificity for a particular instrument. In addition, studies should report sufficient data to enable 2×2 tables to be calculated so that full use can be made of the test accuracy data. Finally, in terms of reporting, studies should provide information such as the typical duration of administration and the level of training required to deliver the test. This information would prove useful for the cost-effectiveness analysis.

Limitations of the clinical effectiveness studies

The Cochrane risk of bias tool[57] was used to assess the quality of the clinical effectiveness studies. This also revealed that a majority of items were rated as being either at unclear or high risk of bias. The large number of items rated as unclear suggests that future studies should refer to the Consolidated Standards of Reporting Trials (CONSORT) statement,[94] and the extension of the statement to non-pharmacological treatments for studies of psychological interventions,[95] to guide the reporting of future trials in this area. In particularly, future studies should describe the journey of all participants through the trial using a CONSORT flow chart. There are concerns that the uptake and dropout rate from mental health interventions may be lower and higher, respectively, in this population than in others and so it is important that these figures are accurately reported. This information will also prove useful in informing future cost-effectiveness evaluations.

A number of studies examined an intervention that had a broad aim (e.g. to reduce stress associated with incarceration) and as a result examined a wide range of outcome measures. Future studies should seek to specify a priori which measure is to be considered the primary outcome to protect against the post hoc selection of measures. More generally, the publication of trial protocols in which the primary outcome is stated along with a list of all secondary outcome measures would help to protect against selective reporting bias.

Chapter 12 Conclusion

The previous chapter summarised some of the limitations of the primary literature and the evidence synthesis. Perhaps the main limitation, however, is that the lack of sufficient primary studies make it difficult to answer any of the review objectives with any degree of certainty.

Implications

Current UK policy recommends the use of the mental health component of the CHAT as the screening measure for the identification of mental health problems among young people who offend. Our review identified no trials of the clinical effectiveness and cost-effectiveness of screening in this population. Furthermore, we identified limited evidence relating to the clinical effectiveness of interventions for mental health problems in this group. Many of the trials were conducted in the USA and many were conducted in custodial settings. In contrast, the majority of young people who offend in the UK are managed in community settings. It remains unclear if such interventions are effective in young people who offend, particularly as applied to a UK setting. This is important because screening can be of value only if it can be linked to an effective intervention.

Our review also identified limited data on the diagnostic test accuracy of screening measures in this population. As with the clinical trial data, the majority of the studies were conducted in the USA and many were carried out in settings that may differ from typical settings in the UK for young people who offend. We are aware, however, that diagnostic test accuracy data for the CHAT have recently been published[23] and this may alter this conclusion.

Although we developed a decision model, this is, as described previously, best seen as an exemplar because the data on which the model is based are extremely limited. However, the model is of use in identifying parameters that may act as key drivers determining whether or not screening is likely to be cost-effective; these are of relevance in determining future research recommendations.

Summary of research recommendations

The limited evidence that we identified suggests that there are a number of areas of uncertainty and a need for future research to reduce this uncertainty. Full details of the recommended research priorities were provided in *Chapter 10*; the main points are summarised here.

- In terms of clinical effectiveness, the limitations of the existing randomised controlled trial evidence base suggest that further feasibility trials of clinical effectiveness are needed to establish important parameters ahead of definitive trials. Future trials should gather information to permit the calculation of QALYs and should seek to assess whether or not treatment for mental health problems alters intersectoral outcomes, including recidivism.
- There is a need for validation studies in which the performance of a range of screening measures is directly compared against a gold standard diagnostic interview conducted to internationally recognised criteria. Screening measures currently recommended for use in the UK to identify mental health difficulties among young people who have offended, specifically the mental health screen of the CHAT, should be directly compared against other available screening measures as part of such studies. Studies should seek to calculate the diagnostic performance of measures in identifying previously unknown cases.
- This fundamental work on diagnostic test accuracy and clinical effectiveness should be conducted before a trial of screening in this area.
- Evidence was lacking for both community and incarcerated settings so these recommendations apply equally to both settings.

Acknowledgements

We would like to thank the expert advisory group and stakeholder groups for their help in the production of this report. We would also like to thank all of the primary study authors for answering our requests for further information. In addition, we would like to thank Alice North and Philippa Sallows for their help at various stages in the production of this report.

Contributions of authors

Rachel Richardson (Research Fellow) was responsible for study selection, data extraction and quality assessment.

Dominic Trépel (Health Economist) contributed to all aspects of the economic sections and took primary responsibility for the drafting of the economic chapters.

Amanda Perry (Senior Lecturer in Forensic Psychology) was responsible for study selection, data extraction and quality assessment.

Shehzad Ali (Research Fellow, Health Economist) advised on all aspects of the economic sections.

Steven Duffy (Information Specialist) carried out the literature searches and took primary responsibility for drafting the search strategies section of the report.

Rhian Gabe (Statistician) contributed to all aspects of the statistical analysis.

Simon Gilbody (Professor of Mental Health Services Research) provided advice throughout the project on systematic review methods and contributed to the writing of the report.

Julie Glanville (Information Specialist) advised on all aspects of the development of the search strategy.

Catherine Hewitt (Statistician) advised on all aspects of the statistical analysis.

Laura Manea (Lecturer in Psychiatry) contributed to all aspects of the statistical analysis.

Stephen Palmer (Professor, Health Economist) advised on all aspects of the economic sections.

Barry Wright (Professor of Child and Adolescent Psychiatry) provided clinical advice throughout the project and contributed to the writing of the report.

Dean McMillan (Senior Lecturer in Mental Health Services Research) had overall responsibility for the project and took primary responsibility for the drafting of the report.

All of the authors contributed to and commented on the report.

References

1. National Audit Office. *The Youth Justice System in England and Wales: Reducing Offending in Young People*. HC 663, Session 2010–2011. London: National Audit Office; 2010.
2. Fazel S, Doll H, Långström N. Mental disorders among adolescents in juvenile detention and correctional facilities: a systematic review and metaregression analysis of 25 surveys. *J Am Acad Child Adolesc Psychiatry* 2008;**47**:1010–90. http://dx.doi.org/10.1097/CHI.ObO13e31817eecf3
3. Chitsabesan P, Kroll L, Bailey S, Kenning C, Sneider S, MacDonald W, *et al.* Mental health needs of young offenders in custody and in the community. *Br J Psychiatry* 2006;**188**:534–40. http://dx.doi.org/10.1192/bjp.bp.105.010116
4. Harrington R, Bailey S. *Mental Health Needs and Effectiveness of Provision for Young Offenders in Custody and the Community*. London: Youth Justice Board; 2005.
5. Chitsabesan P, Bailey S, Williams R, Kroll L, Kenning C, Talbot L. Learning disabilities and educational needs of juvenile offenders. *J Child Serv* 2007;**2**:4–16.
6. Remschmidt H, Reinhard W. The long-term outcome of delinquent children: a 30-year follow-up study. *J Neural Transm* 2010;**117**:663–77. http://dx.doi.org/10.1007/s00702-010-0393-8
7. Fergusson DM, Horwood LJ, Ridder EM. Show me the child at seven: the consequences of conduct problems in childhood for psychosocial functioning in adulthood. *J Child Psychol Psychiatry* 2005;**46**:837–49. http://dx.doi.org/10.1111/j.1469-7610.2004.00387.x
8. Sainsburys Centre for Mental Health. *The Chance of a Lifetime. Preventing Early Conduct Problems and Reducing Crime*. London: Sainsburys Centre for Mental Health; 2009.
9. Barrett B, Byford S, Chitsabesan P, Kenning C. Mental health provision for young offenders: service use and cost. *Br J Psychiatry* 2006;**188**:541–6. http://dx.doi.org/10.1192/bjp.bp.105.010108
10. Grisso T, Barnum R. *Massachusetts Youth Screening Instrument – Second Version: User's Manual and Technical Report*. Sarasota, FL: Professional Resource Press; 2003.
11. Mant D, Fowler G. Mass screening: theory and ethics. *BMJ* 1990;**300**:916–18. http://dx.doi.org/10.1136/bmj.300.6729.916
12. Compton SN, March JS, Brent D, Albano AM, Weersing VR, Curry J. Cognitive–behavioral psychotherapy for anxiety and depressive disorders in children and adolescents: an evidence-based medicine review. *J Am Acad Child Adolesc Psychiatry* 2004;**43**:930–59. http://dx.doi.org/10.1097/01.chi.0000127589.57468.bf
13. Reinblatt SP, Riddle MA. The pharmacological management of childhood anxiety disorders: a review. *Psychopharmacology* 2007;**191**:67–86. http://dx.doi.org/10.1007/s00213-006-0644-4
14. Stallard P. Psychological interventions for post-traumatic reactions in children and young people: a review of randomised controlled trials. *Clin Psychol Rev* 2006;**26**:895–911. http://dx.doi.org/10.1016/j.cpr.2005.09.005
15. Walkup JT, Albano AM, Piacentini J, Birmaher B, Compton SN, Sherrill JT, *et al.* Cognitive behavioral therapy, sertraline, or a combination in childhood anxiety. *N Engl J Med* 2008;**359**:2753–66. http://dx.doi.org/10.1056/NEJMoa0804633
16. Van der Oord S, Prins PJM, Oosterlaan J, Emmelkamp PMG. Efficacy of methylphenidate, psychosocial treatments and their combination in school-aged children with ADHD: a meta-analysis. *Clin Psychol Rev* 2008;**28**:783–800. http://dx.doi.org/10.1016/j.cpr.2007.10.007

17. Kryzhanovskaya L, Schulz SC, McDougle C, Frazier J, Dittmann R, Robertson-Plouch C, *et al.* Olanzapine versus placebo in adolescents with schizophrenia: a 6-week, randomized, double-blind, placebo-controlled trial. *J Am Acad Child Adolesc Psychiatry* 2009;**48**:60–70. http://dx.doi.org/10.1097/CHI.0b013e3181900404

18. Townsend E, Walker D-M, Sergeant S, Vostanis P, Hawton K, Stocker O, *et al.* Systematic review and meta-analysis of interventions relevant for young offenders with mood disorders, anxiety disorders, or self-harm. *J Adolesc* 2010;**33**:9–20. http://dx.doi.org/10.1016/j.adolescence.2009.05.015

19. Youth Justice Board. *National Standards for Youth Justice Services*. London: Youth Justice Board; 2004.

20. The Bradley Report. *Lord Bradley's Review of People with Mental Health Problems or Learning Disabilities in the Criminal Justice System*. London: Department of Health; 2009.

21. Department of Health. *Best Practice in Managing Risk*. London: Department of Health; 2007.

22. Mental Health Framework. *To Improve Mental Health and Wellbeing. No Health without Mental Health: Implementation Framework*. London: Department of Health; 2012.

23. Chitsabesan P, Lennox C, Theodosiou L, Law H, Bailey S, Shaw J. The development of the comprehensive health assessment tool for young offenders within the secure estate. *J Forens Psychiatry Psychol* 2014;**25**:1–25. http://dx.doi.org/10.1080/14789949.2014.882387

24. Youth Justice Board. *ASSET: A Summary of the Evaluation of the Validity and Reliability of the Youth Justice Board Assessment for Young Offenders*. London: Youth Justice Board; nd.

25. Youth Justice Board. *The Mental Health Screening Interview for Adolescents (SIfA): Young Person's Interview*. London: Youth Justice Board; 2003.

26. World Health Organization. *The ICD-10 Classification of Mental and Behavioural Disorders*. Geneva: World Health Organization; 1992.

27. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. Washington, DC: American Psychiatric Association; 2000.

28. Bailey S, Tarbuck P. Recent advances in the development of screening tools for mental health in young offenders. *Curr Opin Psychiatry* 2006;**19**:373–7. http://dx.doi.org/10.1097/01.yco.0000228756.72366.de

29. Centre for Reviews and Dissemination. *Systematic Reviews: CRD's Guidance for Undertaking Reviews in Healthcare*. York: University of York; 2009.

30. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *J Clin Epidemiol* 2009;**62**:1006–12. http://dx.doi.org/10.1016/j.jclinepi.2009.06.005

31. Whiting P, Westwood M, Beynon R, Burke M, Sterne JAC, Glanville J. Inclusion of methodological filters in searches for diagnostic test accuracy studies misses relevant studies. *J Clin Epidemiol* 2011;**64**:602–7. http://dx.doi.org/10.1016/j.jclinepi.2010.07.006

32. Hogan TP. *Psychological Testing: A Practical Introduction*. Hoboken, NJ: John Wiley; 2007.

33. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, *et al.* QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;**155**:529–36. http://dx.doi.org/10.7326/0003-4819-155-8-201110180-00009

34. Mann R, Hewitt CE, Gilbody SM. Assessing the quality of diagnostic studies using psychometric instruments: applying QUADAS. *Soc Psychiatry Psychiatr Epidemiol* 2009;**44**:300–7. http://dx.doi.org/10.1007/s00127-008-0440-z

35. Cashel M, Rogers R, Sewell KW, Holliman NB. Clinical correlates of the Minnesota Multiphasic Personality Inventory (MMPI-A) for a male delinquent population. *J Pers Assess* 1998;**71**:49–69. http://dx.doi.org/10.1207/s15327752jpa7101_4

36. Grubin D, Carson D, Parsons S. *Report on New Prison Reception Health Screening Arrangements: the Results of a Pilot Study in 10 Prisons*. Newcastle: University of Newcastle, Department of Forensic Psychiatry; 2002.

37. Haapanen R, Steiner H. *Identifying Mental Health Treatment Needs among Serious Institutionalized Delinquents using Paper-and Pencil Screening Instruments*. Washington, DC: National Institute of Justice; 2003.

38. Hayes MA, McReynolds LS, Wasserman GA. Paper and voice MAYSI-2: format comparability and concordance with the voice DISC-IV. *Assessment* 2005;**12**:395–403. http://dx.doi.org/10.1177/1073191105280359

39. Kerig P, Arnzen Moeddel M, Becker S. Assessing the sensitivity and specificity of the MAYSI-2 for detecting trauma among youth in juvenile detention. *Child Youth Care Forum* 2011;**40**:345–62. http://dx.doi.org/10.1007/s10566-010-9124-4

40. Kuo ES, Vander Stoep A, Stewart DG. Using the Short Mood and Feelings Questionnaire to detect depression in detained adolescents. *Assessment* 2005;**12**:374–83. http://dx.doi.org/10.1177/1073191105279984

41. McReynolds LS, Wasserman GA, Fisher P, Lucas CP. Diagnostic screening with incarcerated youths: comparing the DPS and Voice DISC. *Crim Justice Behav* 2007;**34**:830–45. http://dx.doi.org/10.1177/0093854807299918

42. Vreugdenhil C, van den Brink W, Ferdinand R, Wouters L, Doreleijers T. The ability of YSR scales to predict DSM/DISC-C psychiatric disorders among incarcerated male adolescents. *Eur Child Adolesc Psychiatry* 2006;**15**:88–96. http://dx.doi.org/10.1007/s00787-006-0497-8

43. Wasserman GA, McReynolds LS, Ko SJ, Katz LM, Cauffman E, Haxton W, *et al.* Screening for emergent risk and service needs among incarcerated youth: comparing MAYSI-2 and voice DISC-IV. *J Am Acad Child Adolesc Psychiatry* 2004;**43**:629–39. http://dx.doi.org/10.1097/00004583-200405000-00017

44. Wasserman GA, Vilhauer JS, McReynolds L, Shoai R, Jonh R. Mental health screening in the juvenile justic system: a comparison between the Voice DISC-IV and the MAYSI-2. *J Juv Justice Serv* 2004;**19**:7–17.

45. Arnzen Moeddel M. *Investigating the Sensitivity of the MAYSI-2 for Detecting PTSD among Female and Male Delinquents*. Master's thesis. Oxford, OH: Miami University; 2008.

46. Angold A, Cosetllo EJ, Loeber R, Messer SC, Pickles A, Winder F, *et al.* Development of a short questionnaire for use in epidemiological studies of depression in children and adolescents: factor composition and structure across development. *Int J Methods Psychiatric Res* 1995;**5**:251–62.

47. Lucas CP, Zhang H, Fisher PW, Shaffer D, Regier DA, Narrow WE, *et al.* The DISC predictive scales (DPS): efficently screening for diagnoses. *J Am Acad Child Adolesc Psychiatry* 2001;**40**:443–9. http://dx.doi.org/10.1097/00004583-200104000-00013

48. Shaffer D, Restifo K, Lucas CP, Dulcan MK, Schwab-Stone ME. NIMH Diagnostic Interview Schedule for Children version IV (NIHM DISC-IV): description, differences from previous versions, and reliability of some common diagnoses. *J Am Acad Child Adolesc Psychiatry* 2000;**39**:29–38. http://dx.doi.org/10.1097/00004583-200001000-00014

49. Archer RP. *MMPI-A: Assessing Adolescent Psychopathology*. 2nd edn. Hillsdale, NJ: Lawrence Erlbaum Associates; 1992.

50. Achenbach TM. *Integrative Guide for the 1991 CBCL/4–18, YSR and TRF Profiles*. Burlington, VT: University of Vermont, Department of Psychiatry; 1991.

51. Endicott J, Spitzer RL. A diagnostic interview: the schedule for affective disorders and schizophrenia. *Arch Gen Psychiatry* 1978;**35**:837–44. http://dx.doi.org/10.1001/archpsyc.1978.01770310043002

52. Pynoos R, Rodriguez N, Steinberg A, Stuber M. *UCLA PTSD Index for DSM-IV (Revision 1)*. Los Angeles, CA: University of California at Los Angeles; 1998.

53. Ambrosini PJ. *Schedule for Affective Disorders and Schizophrenia for School Age Children (6–18 Years): Kiddie-SADS (K-SADS) (Present State Version)*. Philadelphia, PA: Philadelphia Medical College of Pennsylvania; 1992.

54. Baker K, Jones S, Roberts C, Merrington S. *The Evaluation of the Validity and Reliability of the Youth Justice Board's Assessment for Young Offenders: Findings From the First Two Years of the Use of ASSET*. London: Youth Justice Board; 2003.

55. Kroll L, Woodham A, Rothwell J, Bailey S, Tobias C, Harrington R, *et al.* Reliability of the Salford needs assessment schedule for adolescents. *Psychol Med* 1999;**29**:891–902. http://dx.doi.org/10.1017/S0033291799008752

56. Mitchell P, Smedley K, Kenning C, McKee A, Woods D, Rennie C, *et al.* Cognitive–behavioural therapy for adolescents with mental health problems in custody. *J Adolesc* 2011;**34**:433–43. http://dx.doi.org/10.1016/j.adolescence.2010.06.009

57. Higgins JPT, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, *et al.* The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011;**343**:d5928. http://dx.doi.org/10.1136/bmj.d5928

58. Higgins JPT, Green S. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0*. 2011. URL: www.cochrane-handbook.org (accessed 30 April 2013).

59. Ahrens J, Rexford L. Cognitive processing therapy for incarcerated adolescents with post traumatic stress disorder. *J Aggress Maltreat Trauma* 2002;**6**:201–16. http://dx.doi.org/10.1300/J146v06n01_10

60. Biggam FH, Power KG. A controlled, problem-solving, group-based intervention with vulnerable incarcerated young offenders. *Int J Offender Ther Comp Criminol* 2002;**46**:678–98. http://dx.doi.org/10.1177/0306624X02238162

61. Martsch MD. A comparison of two group interventions for adolescent aggression: high process versus low process. *Res Soc Work Pract* 2005;**15**:8–18. http://dx.doi.org/10.1177/1049731504267333

62. Persons R. Psychological and behavioral change in delinquents following psychotherapy. *J Clin Psychol* 1966;**22**:337–40. http://dx.doi.org/10.1002/1097-4679(196607)22:3<337::AID-JCLP2270220328>3.0.CO;2-0

63. Reardon J. *The Effects of Rational Stage Directed Therapy on Self Concept and Psychological Stress in Adolescent Delinquent Females*. PhD thesis. Columbus, OH: Ohio State University; 1976.

64. Rohde P, Clarke GN, Mace DE, Jorgensen JS, Seeley JR. An efficacy/effectiveness study of cognitive–behavioral treatment for adolescents with comorbid major depression and conduct disorder. *J Am Acad Child Adolesc Psychiatry* 2004;**43**:660–8. http://dx.doi.org/10.1097/01.chi.0000121067.29744.41

65. Rohde P, Jorgensen JS, Seeley JR, Mace DE. Pilot evaluation of the coping course: a cognitive–behavioral intervention to enhance coping skills in incarcerated youth. *J Am Acad Child Adolesc Psychiatry* 2004;**43**:669–76. http://dx.doi.org/10.1097/01.chi.0000121068.29744.a5

66. Scherer D, Brondino M, Henggeler S, Melton G, Hanley J. Multisystemic family preservation therapy: preliminary findings from a study of rural and minority serious adolescent offenders. *J Emot Behav Disord* 1994;**2**:198–206. http://dx.doi.org/10.1177/106342669400200402

67. Shivrattan JL. Social interactional training and incarcerated juvenile delinquents. *Can J Criminol* 1988;**30**:145–63.

68. D'Zurilla TJ, Goldfried MR. Problem solving and behavior modification. *J Abnorm Psychol* 1971;**78**:107–26. http://dx.doi.org/10.1037/h0031360

69. Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. *Lancet* 2002;**359**:614–18. http://dx.doi.org/10.1016/S0140-6736(02)07750-4

70. Beck AT, Steer RA, Brown GK. *Manual for Beck Depression Inventory-II*. San Antonio, TX: Psychological Corporation; 1996.

71. Hamilton M. A rating scale for depression. *J Neurol Neurosurg* 1960;**23**:56–61. http://dx.doi.org/10.1136/jnnp.23.1.56

72. Horowitz M, Wilner N, Alvarez W. Impact of events scale: a measure of subjective distress. *Psychosom Med* 1979;**41**:421–3. http://dx.doi.org/10.1097/00006842-197905000-00004

73. Foa E, Riggs DS, Dancu CV, Rothbaum BO. Reliability and validity of a brief instrument for assessing post-traumatic stress disorder. *J Trauma Stress* 1993;**6**:459–73. http://dx.doi.org/10.1002/jts.2490060405

74. Beck AT, Ward CH, Mendelsohn M, Mock J, Erbaugh J. An inventory for measuring depression. *Arch Gen Psychiatry* 1961;**4**:561–71. http://dx.doi.org/10.1001/archpsyc.1961.01710120031004

75. Quay HC, Peterson DR. *Revised Behavior Problem Checklist: Professional Manual*. Odessa, FL: Psychological Assessment Resources; 1987.

76. Derogatis LR, Melisaratos N. The Brief Symptom Inventory: an introductory report. *Psychol Med* 1983;**13**:595–60. http://dx.doi.org/10.1017/S0033291700048017

77. Centre for Reviews and Dissemination. *Making Cost-Effectiveness Information Available: The NHS Economic Evaluation Database Project*. York: University of York; 1996.

78. Drummond M. *Methods for the Economic Evaluation of Health Care Programmes*. Oxford: Oxford University Press; 1997.

79. Ministry of Justice. *Mental Health in Youth Justice*. 2011 (updated 13 March 2011). URL: www.justice.gov.uk/youth-justice/health/mental-health (accessed 30 April 2013).

80. Lynch FL, Dickerson JF, Clarke G, Vitiello B, Porta G, Wagner KD, *et al.* Incremental cost-effectiveness of combined therapy vs. medication only for youth with selective serotonin reuptake inhibitor-resistant depression: treatment of SSRI-resistant depression in adolescents trial findings. *Arch Gen Psychiatry* 2011;**68**:253–62. http://dx.doi.org/10.1001/archgenpsychiatry.2011.9

81. Revicki DA, Wood M. Patient-assigned health state utilities for depression-related outcomes: differences by depression severity and antidepressant medications. *J Affect Disord* 1998;**48**:25–36. http://dx.doi.org/10.1016/S0165-0327(97)00117-1

82. Simon GE, Katon WJ, Lin EHB, Rutter C, Manning WG, Von Korff M, *et al.* Cost-effectiveness of systematic depression treatment among people with diabetes mellitus. *Arch Gen Psychiatry* 2007;**64**:65–72. http://dx.doi.org/10.1001/archpsyc.64.1.65

83. Harshbarger JL. The impact of mental health dimensions on the prediction of juvenile reentry recidivism. Wichita, KA: Wichita State University; 2005.

84. Ministry of Justice. *Proven Re-offending Statistics Quarterly Bulletin July 2010 to June 2011 England and Wales*. London: Ministry of Justice; 2013.

85. Lipsey MW, Landenberger NA, Wilson SJ. Effects of cognitive–behavioral programs for criminal offenders. *Campbell Syst Rev* 2007;**6**:1–27.

86. Netten A, Curtis L. *Unit Costs of Health and Social Care 2012*. Kent: University of Kent, Personal Social Services Research Unit; 2012.

87. Brookes N, Barrett B, Netten A, Knapp E. *Unit Costs in Criminal Justice (UCCJ)*. Kent: University of Kent, Personal Social Services Research Unit; 2013.

88. Brand S, Price R. *The Economic and Social Costs of Crime*. Home Office Research Study 217. London: Development and Statistics Directorate, Home Office; 2000.

89. Claxton K, Walker S, Palmer S, Sculpher M. *Appropriate Perspectives for Health Care Decisions*. York: University of York; 2010.

90. Beck AT, Steer RA, Carbin MG. Psychometric properties of the Beck Depression Inventory: twenty-five years of evaluation. *Clin Psychol Rev* 1988;**8**:77–100. http://dx.doi.org/10.1016/0272-7358(88)90050-5

91. Birmingham L, Mason D, Grubin D. Health screening at first reception into prison. *J Forensic Psychiatry* 1997;**8**:435–9. http://dx.doi.org/10.1080/09585189708412022

92. Byford S, Barrett D, Roberts C, Wilkinson P, Dubicka B, Kelvin RG, *et al.* Cost-effectiveness of selective serotonin reuptake inhibitors and routine specialist care with and without cognitive–behavioural therapy in adolescents with major depression. *Br J Psychiatry* 2007;**191**:521–7. http://dx.doi.org/10.1192/bjp.bp.107.038984

93. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CE, Glasziou PP, Irwig LM, *et al.* Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ* 2003;**326**:41–4. http://dx.doi.org/10.1136/bmj.326.7379.41

94. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. *Ann Intern Med* 2010;**152**:726–32. http://dx.doi.org/10.7326/0003-4819-152-11-201006010-00232

95. Boutron I, Moher D, Altman GD, Schulz KF, Ravaud P. Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: explanation and elaboration. *Ann Intern Med* 2008;**148**:295–309. http://dx.doi.org/10.7326/0003-4819-148-4-200802190-00008

Appendix 1 Methods of assessing diagnostic test accuracy

There are standard strategies for assessing the diagnostic test accuracy of screening measures. These involve examining the level of agreement between the screening measure and a recognised 'gold standard' method of establishing a diagnosis. A convenient way of summarising the level of agreement is to use a 2 × 2 table, as shown in *Table 35*.

The four cells (a, b, c, d) capture the four possible relationships between the results of a screening test and a gold standard diagnosis. Screening for depression is used here as an example, but the same basic principles apply to the evaluation of a screening instrument for any diagnosis:

- true positive (cell a): when the person scores positive on the screening test and does in fact have the condition (e.g. the person scores positive for depression on a depression screening measure and meets criteria for major depression according to the gold standard diagnosis)
- false positive (cell b): when the person scores positive on the screening test but does not in fact have the condition (e.g. the person scores positive for depression on a depression screening measure but does not meet criteria for major depression according to the gold standard diagnosis)
- false negative (cell c): when the person scores negative on the screening test but does in fact have the condition (e.g. the person scores as not depressed on a depression screening measure but does meet criteria for major depression according to the gold standard diagnosis)
- true negative (cell d): when the person scores negative on the screening test and does not in fact have the condition (e.g. the person scores as not depressed on a screening test and does not meet criteria for major depression according to the gold standard).

If the numbers of people in a study who are classified as true positives, false positives, false negatives and true negatives are entered into each of the cells, it is possible to calculate a number of indicators of how well the test performs. Two of the most commonly used indicators are sensitivity and specificity. Sensitivity is the proportion of people with the diagnosis who score positive on the screening instrument. In terms of the cells in the 2 × 2 table, this is calculated as a/(a + c). Specificity refers to the proportion of people without the diagnosis who score negative on the screening instrument and is calculated as d/(b + d).

Often a screening measure will have a range of scores. For example, even brief measures of depression may have possible scores ranging from 0 to ≥ 30. Different cut-off points could therefore be used to classify someone as being positive for depression on that measure (e.g. anyone scoring ≥ 1 is classified as positive, anyone scoring ≥ 2 is classified as positive). For each of these cut-off points it would be possible to calculate separate 2 × 2 tables, which would lead to separate sensitivity and specificity estimates at each cut-off.

As the cut-off is varied, sensitivity and specificity will change in a constant way: as sensitivity increases, specificity will decrease (and vice versa). There is, then, always a balance to be struck: if sensitivity is high, specificity is likely to be low; if specificity is high, sensitivity is likely to be low. A decision needs to be made

TABLE 35 A 2 × 2 table for summarising test accuracy

	Gold standard diagnosis	
Screening measure	+	–
+	True positive (a)	False positive (b)
–	False negative (c)	True negative (d)

about what balance between sensitivity and specificity is likely to be appropriate in a particular decision-making situation or clinical context.

There are a number of other commonly used methods for estimating the diagnostic accuracy of a screening method in addition to sensitivity and specificity. Descriptions of these additional statistics are given below:

- positive predictive value: the proportion of people who score positive on the screening instrument who have the diagnosis, which in terms of the 2 × 2 table above is calculated as a/(a + b)
- negative predictive value: the proportion of people who score negative on the screening instrument who do not have the diagnosis, calculated as d/(c + d)
- positive likelihood ratio: the probability of a person who meets criteria for the diagnosis testing positive divided by the probability of a person who does not meet criteria for the diagnosis testing positive, calculated as sensitivity/(1 – specificity)
- negative likelihood ratio: the probability of a person who meets criteria for the diagnosis testing negative divided by the probability of a person who does not meet criteria for the diagnosis testing negative, calculated as (1 – sensitivity)/specificity
- DOR: the odds of a screening measure being positive if a person meets criteria for the diagnosis relative to the odds of the screening measure being positive if the person does not meet criteria for the diagnosis, calculated as positive likelihood ratio/negative likelihood ratio.

Appendix 2 Stakeholders and advisors

Sarah Byford	Institute of Psychiatry, London
Saima Bouden	Junior Youth Inclusion Project, Nacro, Leeds
Louise Dare	Care and Programmes Manager, Aycliffe Secure Services, Newton Aycliffe, County Durham
David Edwards	Wandsworth Children and Adolescent Mental Health Services/Young Offender Team, South West London and St George's Mental Health NHS Trust
Don Grubin	Professor of Forensic Psychiatry, University of Newcastle
Alison Eastwood	Senior Review Manager, Centre for Reviews and Dissemination, University of York
Howard Jasper	Strategy Manager (Health and Accommodation), Youth Justice Board for England and Wales, London
Tracey McGowan	Primary Care Mental Health, Her Majesty's Young Offenders Institution, Wetherby
Emma Palmer	Reader in Forensic Psychology, University of Leicester

Appendix 3 Database searches

Summary of database searches

Resource	No. of papers identified
PsycINFO	5915
MEDLINE	4429
EMBASE	1917
CDSR	45
DARE	25
CENTRAL	373
HTA database	0
NHS EED	21
ASSIA	704
Criminal Justice Abstracts	1706
NCJRS	959
Social Policy & Practice	1089
Social Services Abstracts	524
PAIS International	106
SCI	483
SSCI	1531
CPCI-S	56
CPCI-SSH	134
Social Care Online	635
Campbell Library	89
HEED	5
OAIster	18
Index of Theses	5
Zetoc	9
RePEc	1
Internet: organisation website searches	97
Total	20,876
Total after deduplication	13,527

Search strategies for electronic databases

PsycINFO (OvidSP), 1806 to 2011 March Week 4

Searched 5 April 2011.

1. (adolescence 13 17 yrs or young adulthood 18 29 yrs).ag. (434,650)
2. (adolescen$ or juvenile$ or youth$ or teenage$ or youngster$).ti,ab. (178,061)
3. (young people or young person or young persons or young adult$ or early adult$).ti,ab. (35,630)
4. or/1-3 (501,937)
5. crime/ (10,213)
6. exp criminals/ (13,825)
7. prisoners/ (7337)
8. prisons/ (4033)
9. or/5-8 (29,096)
10. 4 and 9 (6540)
11. (secure adj2 (placement or accommodation or facilit$ or care or unit$ or centre$ or center$ or home$ or setting$)).ti,ab. (789)
12. high dependency unit$.ti,ab. (16)
13. 4 and (11 or 12) (285)
14. exp juvenile delinquency/ (14,243)
15. (young adj2 offend$).ti,ab. (1017)
16. ((adolescen$ or juvenile$ or youth$ or teenage$ or youngster$) adj3 (criminal$ or crime$ or penal or justice or custody or custodi$ or probation or parole$ or convict$ or reconvict$ or incarcerat$ or judicial$ or justice or sentence$ or court or remand$)).ti,ab. (5499)
17. ((young people or young person or young persons or young adult$ or early adult$) adj3 (criminal$ or crime$ or penal or justice or custody or custodi$ or probation or parole$ or convict$ or reconvict$ or incarcerat$ or judicial$ or justice or sentence$ or court or remand$)).ti,ab. (211)
18. ((kid or kids or boy or boys or girl or girls or child or children) adj3 (criminal$ or crime$ or penal or justice or custody or custodi$ or probation or parole$ or convict$ or reconvict$ or incarcerat$ or judicial$ or justice or sentence$)).ti,ab. (3355)
19. ((adolescen$ or juvenile$ or youth$ or teenage$ or youngster$) adj3 (offend$ or offence$ or reoffend $ or reoffence$ or delinquen$ or court or remand$)).ti,ab. (10,111)
20. ((young people or young person or young persons or young adult$ or early adult$) adj3 (offend$ or offence$ or reoffend$ or reoffence$ or delinquen$)).ti,ab. (171)
21. ((kid or kids or boy or boys or girl or girls or child or children) adj3 (offend$ or offence$ or reoffend$ or reoffence$ or delinquen$)).ti,ab. (2770)
22. ((adolescen$ or juvenile$ or youth$ or teenage$ or youngster$) adj3 (prison$ or jail$ or gaol$ or inmate$ or reformator$)).ti,ab. (267)
23. ((young people or young person or young persons or young adult$ or early adult$) adj3 (prison$ or jail$ or gaol$ or inmate$ or reformator$)).ti,ab. (29)
24. ((kid or kids or boy or boys or girl or girls or child or children) adj3 (prison$ or jail$ or gaol$ or inmate$ or reformator$)).ti,ab. (195)
25. or/13-24 (23,751)
26. 10 or 25 (27,555)
27. exp mental health/ (30,180)
28. mental disorders/ (53,316)
29. (mental$ adj (health or disorder$ or disease$ or illness or problem$)).ti,ab. (130,543)
30. Autism/ (15,807)
31. Aspergers Syndrome/ (1749)
32. Pervasive Developmental Disorders/ (3809)
33. Attention Deficit Disorder with Hyperactivity/ (9782)
34. Oppositional Defiant Disorder/ (959)
35. Conduct Disorder/ (2876)

36. Hyperkinesis/ (6758)
37. exp Schizophrenia/ (62,108)
38. Psychosis/ (15,920)
39. exp Affective Disorders/ (96,199)
40. exp Anxiety Disorders/ (49,944)
41. exp Self Destructive Behavior/ (25,600)
42. exp Somatoform Disorders/ (9581)
43. exp Eating Disorders/ (19,169)
44. exp Impulse Control Disorders/ (517)
45. Kleptomania/ (152)
46. Pyromania/ (77)
47. exp Neurosis/ (7100)
48. (autistic or autism or kanner$ syndrome$).ti,ab. (9396)
49. asperger$.ti,ab. (2186)
50. (pervasive development$ adj2 disorder$).ti,ab. (1933)
51. (attention deficit$ or minimal$ brain damage$ or minimal$ brain dysfunction$ or hyperkinetic$ or ADHD or addh or ad hd or hkd).ti,ab. (19,390)
52. (oppositional defian$ disorder$ or ODD or oppositional defian$ problem$).ti,ab. (3676)
53. (disruptive behavio?r$ disorder$ or disruptive behavio?r$ problem$).ti,ab. (1098)
54. (conduct adj2 disorder$).ti,ab. (5023)
55. (hyperkinesis or hyperkinesia$ or hyperkinetic disorder$).ti,ab. (760)
56. ((motor or movement) adj2 hyperactivity).ti,ab. (138)
57. mixed disorder$.ti,ab. (64)
58. (schizophren$ or dementia praecox).ti,ab. (82,891)
59. ((schizoaffective or schizophreniform) adj2 disorder$).ti,ab. (3744)
60. (psychosis or psychoses or psychotic).ti,ab. (47,610)
61. ((mood or affective) adj2 disorder$).ti,ab. (20,551)
62. (depressive or depression$).ti,ab. (158,658)
63. melancholia$.ti,ab. (1834)
64. (dysthymic adj2 disorder$).ti,ab. (894)
65. (bipolar$ adj3 (disorder$ or depress$ or illness$ or disease$ or episod$)).ti,ab. (16,542)
66. (mania or manic).ti,ab. (13,771)
67. (hypomanic or hypo-manic or hypomania or hypo-mania).ti,ab. (2158)
68. cyclothym$.ti,ab. (865)
69. (anxiety adj3 (disorder$ or neurosis or neuroses or neurotic$)).ti,ab. (21,504)
70. (panic adj2 (disorder$ or attack$)).ti,ab. (9502)
71. (phobia$ or phobic$).ti,ab. (11,728)
72. (agoraphobia$ or agoraphobic$).ti,ab. (3953)
73. obsessive compulsive.ti,ab. (11,629)
74. ((stress or traumatic or posttraumatic or post-traumatic or combat) adj2 disorder$).ti,ab. (17,797)
75. (anankastic adj2 personalit$).ti,ab. (15)
76. ((self adj2 (harm$ or injur$ or mutilat$ or poison$ or wound$ or destruct$)) or selfharm).ti,ab. (8965)
77. (suicide$ or suicidal or parasuicid$).ti,ab. (35,231)
78. (delusional adj2 disorder$).ti,ab. (757)
79. ((somati?ati$ or somatoform or briquet or pain) adj2 (disorder$ or syndrom$)).ti,ab. (4596)
80. (conversion adj2 (disorder$ or reaction$ or hysteria$)).ti,ab. (1051)
81. (astasia abasia or globus hystericus).ti,ab. (51)
82. hypochondria$.ti,ab. (2650)
83. (body adj3 image adj3 (disorder$ or d?sfunction$)).ti,ab. (538)
84. (body adj3 dysmorphic).ti,ab. (655)
85. ((eating or appetite) adj2 disorder$).ti,ab. (14,212)
86. anorexi$.ti,ab. (10,789)
87. bulimi$.ti,ab. (8190)

88. (impulse adj2 disorder$).ti,ab. (753)
89. (intermittent adj2 disorder$).ti,ab. (248)
90. kleptomani$.ti,ab. (296)
91. (fireset$ or firestart$ or (fire adj1 set$) or (fire adj1 start$) or arson$ or pyromania$).ti,ab. (786)
92. ((neurotic adj1 disorder$) or neuroses or psychoneuros$).ti,ab. (6699)
93. or/27-92 (529,072)
94. 26 and 93 (5365)
95. mentally ill offenders/ (2786)
96. forensic psychiatry/ (2932)
97. forensic psychology/ (2726)
98. 4 and (95 or 96 or 97) (1010)
99. 94 or 98 (5947)
100. (animal or animals or rat or rats or mouse or mice or hamster or hamsters or dog or dogs or cat or cats or bovine or sheep or ovine or pig or pigs).ab,ti,id,de. (218,806)
101. 99 not 100 (5915)

MEDLINE and MEDLINE In-Process and Other Non-Indexed Citations (OvidSP), 1948 to 2011 March Week 4

Searched 5 April 2011.

1. Adolescent/ (1,390,340)
2. Young Adult/ (135,506)
3. (adolescen$ or juvenile$ or youth$ or teenage$ or youngster$).ti,ab. (196,069)
4. (young people or young person or young persons or young adult$ or early adult$).ti,ab. (57,247)
5. or/1-4 (1,533,916)
6. Crime/ (11,300)
7. Criminals/ (226)
8. Prisoners/ (10,337)
9. Prisons/ (6301)
10. or/6-9 (25,385)
11. 5 and 10 (5213)
12. (secure adj2 (placement or accommodation or facilit$ or care or unit$ or centre$ or center$ or home$ or setting$)).ti,ab. (486)
13. high dependency unit$.ti,ab. (285)
14. 5 and (12 or 13) (147)
15. Juvenile Delinquency/ (6496)
16. (young adj2 offend$).ti,ab. (299)
17. ((adolescen$ or juvenile$ or youth$ or teenage$ or youngster$) adj3 (criminal$ or crime$ or penal or justice or custody or custodi$ or probation or parole$ or convict$ or reconvict$ or incarcerat$ or judicial$ or justice or sentence$ or court or remand$)).ti,ab. (1428)
18. ((young people or young person or young persons or young adult$ or early adult$) adj3 (criminal$ or crime$ or penal or justice or custody or custodi$ or probation or parole$ or convict$ or reconvict$ or incarcerat$ or judicial$ or justice or sentence$ or court or remand$)).ti,ab. (64)
19. ((kid or kids or boy or boys or girl or girls or child or children) adj3 (criminal$ or crime$ or penal or justice or custody or custodi$ or probation or parole$ or convict$ or reconvict$ or incarcerat$ or judicial$ or justice or sentence$ or court or remand$)).ti,ab. (1171)
20. ((adolescen$ or juvenile$ or youth$ or teenage$ or youngster$) adj3 (offend$ or offence$ or reoffend$ or reoffence$ or delinquen$)).ti,ab. (2258)
21. ((young people or young person or young persons or young adult$ or early adult$) adj3 (offend$ or offence$ or reoffend$ or reoffence$ or delinquen$)).ti,ab. (50)
22. ((kid or kids or boy or boys or girl or girls or child or children) adj3 (offend$ or offence$ or reoffend$ or reoffence$ or delinquen$)).ti,ab. (677)

23. ((adolescen$ or juvenile$ or youth$ or teenage$ or youngster$) adj3 (prison$ or jail$ or gaol$ or inmate$ or reformator$)).ti,ab. (109)
24. ((young people or young person or young persons or young adult$ or early adult$) adj3 (prison$ or jail$ or gaol$ or inmate$ or reformator$)).ti,ab. (16)
25. ((kid or kids or boy or boys or girl or girls or child or children) adj3 (prison$ or jail$ or gaol$ or inmate$ or reformator$)).ti,ab. (68)
26. or/14-25 (9069)
27. 11 or 26 (12,695)
28. Mental Health/ (16,560)
29. Mental Disorders/ (105,198)
30. (mental$ adj (health or disorder$ or disease$ or illness or problem$)).ti,ab. (77,054)
31. Autistic Disorder/ (13,219)
32. Asperger Syndrome/ (1106)
33. Child Development Disorders, Pervasive/ (1796)
34. Attention Deficit Disorder with Hyperactivity/ (15,547)
35. "Attention Deficit and Disruptive Behavior Disorders"/ (1513)
36. Conduct Disorder/ (1624)
37. Hyperkinesis/ (3348)
38. exp Schizophrenia/ (73,961)
39. Psychotic Disorders/ (27,402)
40. exp Mood Disorders/ (96,556)
41. exp Anxiety Disorders/ (55,054)
42. exp Self-Injurious Behavior/ (46,441)
43. exp Somatoform Disorders/ (11,396)
44. exp Eating disorders/ (18,704)
45. Impulse Control Disorders/ or Firesetting Behavior/ (1925)
46. Neurotic Disorders/ (14,814)
47. (autistic or autism or kanner$ syndrome$).ti,ab. (6135)
48. asperger$.ti,ab. (1201)
49. (pervasive development$ adj2 disorder$).ti,ab. (1301)
50. (attention deficit$ or minimal$ brain damage$ or minimal$ brain dysfunction$ or hyperkinetic$ or ADHD or addh or ad hd or hkd).ti,ab. (16,729)
51. (oppositional defian$ disorder$ or ODD or oppositional defian$ problem$).ti,ab. (5261)
52. (disruptive behavio?r$ disorder$ or disruptive behavio?r$ problem$).ti,ab. (650)
53. (conduct adj2 disorder$).ti,ab. (2784)
54. (hyperkinesis or hyperkinesia$ or hyperkinetic disorder$).ti,ab. (1434)
55. ((motor or movement) adj2 hyperactivity).ti,ab. (196)
56. mixed disorder$.ti,ab. (87)
57. (schizophren$ or dementia praecox).ti,ab. (71,761)
58. ((schizoaffective or schizophreniform) adj2 disorder$).ti,ab. (2903)
59. (psychosis or psychoses or psychotic).ti,ab. (36,327)
60. ((mood or affective) adj2 disorder$).ti,ab. (18,499)
61. (depressive or depression$).ti,ab. (187,549)
62. melancholia$.ti,ab. (1030)
63. (dysthymic adj2 disorder$).ti,ab. (578)
64. (bipolar$ adj3 (disorder$ or depress$ or illness$ or disease$ or episod$)).ti,ab. (14,852)
65. (mania or manic).ti,ab. (10,903)
66. (hypomanic or hypo-manic or hypomania or hypo-mania).ti,ab. (1604)
67. cyclothym$.ti,ab. (575)
68. (anxiety adj3 (disorder$ or neurosis or neuroses or neurotic$)).ti,ab. (15,887)
69. (panic adj2 (disorder$ or attack$)).ti,ab. (7728)
70. (phobia$ or phobic$).ti,ab. (7349)
71. (agoraphobia$ or agoraphobic$).ti,ab. (2468)

72. obsessive compulsive.ti,ab. (8410)
73. ((stress or traumatic or posttraumatic or post-traumatic or combat) adj2 disorder$).ti,ab. (12,087)
74. (anankastic adj2 personalit$).ti,ab. (12)
75. ((self adj2 (harm$ or injur$ or mutilat$ or poison$ or wound$ or destruct$)) or selfharm).ti,ab. (7770)
76. (suicide$ or suicidal or parasuicid$).ti,ab. (40,671)
77. (delusional adj2 disorder$).ti,ab. (554)
78. ((somati?ati$ or somatoform or briquet or pain) adj2 (disorder$ or syndrom$)).ti,ab. (11,829)
79. (conversion adj2 (disorder$ or reaction$ or hysteria$)).ti,ab. (1198)
80. (astasia abasia or globus hystericus).ti,ab. (91)
81. hypochondria$.ti,ab. (2225)
82. (body adj3 image adj3 (disorder$ or d?sfunction$)).ti,ab. (248)
83. (body adj3 dysmorphic).ti,ab. (552)
84. ((eating or appetite) adj2 disorder$).ti,ab. (9313)
85. anorexi$.ti,ab. (20,076)
86. bulimi$.ti,ab. (5419)
87. (impulse adj2 disorder$).ti,ab. (534)
88. (intermittent adj2 disorder$).ti,ab. (213)
89. kleptomani$.ti,ab. (162)
90. (fireset$ or firestart$ or (fire adj1 set$) or (fire adj1 start$) or arson$ or pyromania$).ti,ab. (829)
91. ((neurotic adj1 disorder$) or neuroses or psychoneuros$).ti,ab. (3180)
92. or/28-91 (623,217)
93. 27 and 92 (3575)
94. Forensic Psychiatry/ (7289)
95. 5 and 94 (1078)
96. 93 or 95 (4396)
97. animals/ not (animals/ and humans/) (3,471,083)
98. 96 not 97 (4396)

EMBASE (OvidSP), 1980 to 2011 Week 13

Searched 5 April 2011.

1. exp *adolescent/ (23,676)
2. (adolescen$ or juvenile$ or youth$ or teenage$ or youngster$).ti,ab. (234,544)
3. (young people or young person or young persons or young adult$ or early adult$).ti,ab. (67,473)
4. or/1-3 (300,630)
5. *offender/ (1426)
6. *crime/ (7501)
7. *prisoner/ (5434)
8. *prison/ (4840)
9. or/5-8 (18,126)
10. 4 and 9 (873)
11. (secure adj2 (placement or accommodation or facilit$ or care or unit$ or centre$ or center$ or home$ or setting$)).ti,ab. (729)
12. high dependency unit$.ti,ab. (398)
13. 4 and (11 or 12) (79)
14. *juvenile delinquency/ (4689)
15. (young adj2 offend$).ti,ab. (419)
16. ((adolescen$ or juvenile$ or youth$ or teenage$ or youngster$) adj3 (criminal$ or crime$ or penal or justice or custody or custodi$ or probation or parole$ or convict$ or reconvict$ or incarcerat$ or judicial$ or justice or sentence$ or court or remand$)).ti,ab. (1746)
17. ((young people or young person or young persons or young adult$ or early adult$) adj3 (criminal$ or crime$ or penal or justice or custody or custodi$ or probation or parole$ or convict$ or reconvict$ or incarcerat$ or judicial$ or justice or sentence$ or court or remand$)).ti,ab. (95)

18. ((kid or kids or boy or boys or girl or girls or child or children) adj3 (criminal$ or crime$ or penal or justice or custody or custodi$ or probation or parole$ or convict$ or reconvict$ or incarcerat$ or judicial$ or justice or sentence$ or court or remand$)).ti,ab. (1331)
19. ((adolescen$ or juvenile$ or youth$ or teenage$ or youngster$) adj3 (offend$ or offence$ or reoffend$ or reoffence$ or delinquen$)).ti,ab. (2668)
20. ((young people or young person or young persons or young adult$ or early adult$) adj3 (offend$ or offence$ or reoffend$ or reoffence$ or delinquen$)).ti,ab. (73)
21. ((kid or kids or boy or boys or girl or girls or child or children) adj3 (offend$ or offence$ or reoffend$ or reoffence$ or delinquen$)).ti,ab. (784)
22. ((adolescen$ or juvenile$ or youth$ or teenage$ or youngster$) adj3 (prison$ or jail$ or gaol$ or inmate$ or reformator$)).ti,ab. (138)
23. ((young people or young person or young persons or young adult$ or early adult$) adj3 (prison$ or jail $ or gaol$ or inmate$ or reformator$)).ti,ab. (20)
24. ((kid or kids or boy or boys or girl or girls or child or children) adj3 (prison$ or jail$ or gaol$ or inmate$ or reformator$)).ti,ab. (79)
25. or/13-24 (8364)
26. 10 or 25 (8641)
27. *mental health/ (17,223)
28. *mental disease/ (75,041)
29. (mental$ adj (health or disorder$ or disease$ or illness or problem$)).ti,ab. (97,491)
30. *autism/ or *asperger syndrome/ (14,466)
31. *attention deficit disorder/ (15,251)
32. *disruptive behavior/ (247)
33. *conduct disorder/ (951)
34. *hyperkinesia/ (1473)
35. *schizophrenia/ (76,192)
36. *psychosis/ (25,902)
37. *mood disorder/ (4372)
38. *anxiety disorder/ (9639)
39. *automutilation/ (4523)
40. exp *suicidal behavior/ (30,409)
41. *somatoform disorder/ (1843)
42. exp *eating disorder/ (20,432)
43. exp *impulse control disorder/ (3338)
44. *neurosis/ (11,221)
45. (autistic or autism or kanner$ syndrome$).ti,ab. (7658)
46. asperger$.ti,ab. (1561)
47. (pervasive development$ adj2 disorder$).ti,ab. (1705)
48. (attention deficit$ or minimal$ brain damage$ or minimal$ brain dysfunction$ or hyperkinetic$ or ADHD or addh or ad hd or hkd).ti,ab. (21,547)
49. (oppositional defian$ disorder$ or ODD or oppositional defian$ problem$).ti,ab. (6838)
50. (disruptive behavio?r$ disorder$ or disruptive behavio?r$ problem$).ti,ab. (793)
51. (conduct adj2 disorder$).ti,ab. (3489)
52. (hyperkinesis or hyperkinesia$ or hyperkinetic disorder$).ti,ab. (1674)
53. ((motor or movement) adj2 hyperactivity).ti,ab. (247)
54. mixed disorder$.ti,ab. (110)
55. (schizophren$ or dementia praecox).ti,ab. (89,497)
56. ((schizoaffective or schizophreniform) adj2 disorder$).ti,ab. (3794)
57. (psychosis or psychoses or psychotic).ti,ab. (47,149)
58. ((mood or affective) adj2 disorder$).ti,ab. (23,940)
59. (depressive or depression$).ti,ab. (232,725)
60. melancholia$.ti,ab. (1276)
61. (dysthymic adj2 disorder$).ti,ab. (777)

62. (bipolar$ adj3 (disorder$ or depress$ or illness$ or disease$ or episod$)).ti,ab. (20,163)
63. (mania or manic).ti,ab. (13,303)
64. (hypomanic or hypo-manic or hypomania or hypo-mania).ti,ab. (2073)
65. cyclothym$.ti,ab. (701)
66. (anxiety adj3 (disorder$ or neurosis or neuroses or neurotic$)).ti,ab. (21,781)
67. (panic adj2 (disorder$ or attack$)).ti,ab. (10,198)
68. (phobia$ or phobic$).ti,ab. (9574)
69. (agoraphobia$ or agoraphobic$).ti,ab. (3163)
70. obsessive compulsive.ti,ab. (11,289)
71. ((stress or traumatic or posttraumatic or post-traumatic or combat) adj2 disorder$).ti,ab. (15,435)
72. (anankastic adj2 personalit$).ti,ab. (16)
73. ((self adj2 (harm$ or injur$ or mutilat$ or poison$ or wound$ or destruct$)) or selfharm).ti,ab. (9579)
74. (suicide$ or suicidal or parasuicid$).ti,ab. (48,675)
75. (delusional adj2 disorder$).ti,ab. (831)
76. ((somati?ati$ or somatoform or briquet or pain) adj2 (disorder$ or syndrom$)).ti,ab. (16,103)
77. (conversion adj2 (disorder$ or reaction$ or hysteria$)).ti,ab. (1607)
78. (astasia abasia or globus hystericus).ti,ab. (109)
79. hypochondria$.ti,ab. (2716)
80. (body adj3 image adj3 (disorder$ or d?sfunction$)).ti,ab. (321)
81. (body adj3 dysmorphic).ti,ab. (658)
82. ((eating or appetite) adj2 disorder$).ti,ab. (12,944)
83. anorexi$.ti,ab. (24,551)
84. bulimi$.ti,ab. (7376)
85. (impulse adj2 disorder$).ti,ab. (781)
86. (intermittent adj2 disorder$).ti,ab. (269)
87. kleptomani$.ti,ab. (206)
88. (fireset$ or firestart$ or (fire adj1 set$) or (fire adj1 start$) or arson$ or pyromania$).ti,ab. (1120)
89. ((neurotic adj1 disorder$) or neuroses or psychoneuros$).ti,ab. (3553)
90. or/27-89 (671,025)
91. 26 and 90 (1831)
92. *forensic psychiatry/ (6586)
93. 4 and 92 (225)
94. 91 or 93 (2015)
95. Animal/ or Animal Experiment/ or Nonhuman/ (5,478,677)
96. (rat or rats or mouse or mice or murine or rodent or rodents or hamster or hamsters or pig or pigs or porcine or rabbit or rabbits or animal or animals or dogs or dog or cats or cow or bovine or sheep or ovine or monkey or monkeys).ti,ab,sh. (4,270,799)
97. 95 or 96 (5,858,750)
98. exp Human/ or Human Experiment/ (12,314,900)
99. 7 not (97 and 98) (5400)
100. 94 not 99 (1917)

Appendix 4 Excluded studies

TABLE 36 Key to excluded studies

Key	Reason	No. of studies excluded
1	Not a study of a screening tool or an intervention	55
2	Unavailable	7
3A	Not correct population (screening papers)	20
3B	Not correct population (intervention papers)	24
4	Not a psychological or pharmacological intervention	1
5	Not a gold standard comparison	28
6A	Not a relevant outcome (screening papers)	13
6B	Not a relevant outcome (intervention papers)	36
7	Sample overlap	2
8	Not correct study design	14

TABLE 37 List of excluded studies with reasons

No.	Study	Reason
1.	Alexander J, Parsons B. Short-term behavioral intervention with delinquent families: impact on family process and recidivism. *J Abnorm Psychol* 1973;**81**:219–25	6B
2.	Anonymous. Some specific crime costs and prevention savings. *Surveys* 1999	2
3.	Anonymous. The mental health of juvenile offenders. *J Psychosoc Nurs Ment Health Serv* 1999;**37**:9–10	1
4.	Archer RP, Stredny RV, Mason JA, Arnau RC. An examination and replication of the psychometric properties of the Massachusetts Youth Screening Instrument – second edition (MAYSI-2) among adolescents in detention settings. *Assessment* 2004;**11**:290–302	5
5.	Archer RP, Zoby M, Stredny RV. The Minnesota Multiphasic Personality Inventory – Adolescent. In Archer RP, editor. *Forensic Uses of Clinical Assessment Instruments.* Mahwah, NJ: Lawrence Erlbaum; 2006. pp. 57–87	5
6.	Arnzen Moeddel M. *Investigating the Sensitivity of the MAYSI-2 for Detecting PTSD among Female and Male Delinquents*: Oxford, OH: Miami University; 2008	7
7.	Ash EM. *Gender Differences in Psychopathology among Incarcerated Adolescents with a History of Violence.* PhD thesis. Charlottesville, VA: University of Virginia; 1998	5
8.	Bailey SM. *Predicting Mental Illness with the MAYSI-2.* PhD thesis. Minneapolis, MN: Capella University; 2008	6A
9.	Baker K, Jones S, Merrington S, Roberts C. *Youth Justice Board for England and Wales. Further Development of Asset.* London: Youth Justice Board; 2005	6A
10.	Barrett B, Byford S, Chitsabesan P, Kenning C. Mental health provision for young offenders: service use and cost. *Br J Psychiatry* 2006;**188**:541–6	1
11.	Barth RP, Greeson JKP, Guo S, Green RL, Hurley S, Sisson J. Outcomes for youth receiving intensive in-home therapy or residential care: a comparison using propensity scores. *Am J Orthopsychiatry* 2007;**77**:497–505	3B
12.	Beal D, Duckro P. Family counseling as an alternative to legal action for the juvenile status offender. *J Marital Fam Ther* 1977;**3**:77–81	3B

continued

TABLE 37 List of excluded studies with reasons (*continued*)

No.	Study	Reason
13.	Becker VB. *A Comparison of the MMPI and Minimult with Psychotic Delinquents.* PhD thesis. San Diego, CA: United States International University; 1981	5
14.	Biederman J, Faraone SV, Doyle A, Lehman BK, Kraus I, Perrin J, *et al.* Convergence of the Child Behavior Checklist with structured interview-based psychiatric diagnoses of ADHD children with and without comorbidity. *J Child Psychol Psychiatry* 1993;**34**:1241–51	3A
15.	Bittman B, Dickson L, Coddington K. Creative musical expression as a catalyst for quality-of-life improvement in inner-city adolescents placed in a court-referred residential treatment program. *Adv Mind Body Med* 2009;**24**:8–19	2
16.	Borduin C, Mann B, Cone L, Henggeler S, Fucci B, Blaske D, *et al.* Multisystemic treatment of serious juvenile offenders: long term prevention of criminality and violence. *J Consult Clin Psychol* 1995;**63**:569–78	6B
17.	Borduin CM, Henggeler SW, Blaske DM, Stein RJ. Multisystemic treatment of adolescent sexual offenders. *Int J Offender Ther Comp Criminol* 1990;**996**:1	6B
18.	Brannon JM, Williams D. The effectiveness of the Child Behavior Checklist in identifying the behavioral patterns and security requirements of juvenile offenders. *Int J Offender Ther Comp Criminol* 1986;**30**:195–201	5
19.	Burton D, Foy DW, Bwanausi C, Johnson J, Moore L. The relationship between traumatic exposure, family dysfunction, and post-traumatic stress symptoms in male juvenile offenders. *J Trauma Stress* 1994;**7**:83–93	1
20.	Caldwell MF, Vitacco M, Van Rybroek GJ. Are violent delinquents worth treating? A cost–benefit analysis. *J Res Crime Delinq* 2006;**43**:148–68	6B
21.	Cashel ML. *Clinical Correlates of the Minnesota Multiphasic Personality Inventory (MMPI-A) for a Male Delinquent Population.* PhD thesis. Denton, TX: University of North Texas; 1998	1
22.	Cashel ML, Ovaert L, Holliman NG. Evaluating PTSD in incarcerated male juveniles with the MMPI-A: an exploratory analysis. *J Clin Psychol* 2000;**56**:1535–49	5
23.	Cauffman E. A statewide screening of mental health symptoms among juvenile offenders in detention. *J Am Acad Child Adolesc Psychiatry* 2004;**43**:430–9	5
24.	Cauffman E, MacIntosh R. A Rasch differential item functioning analysis of the Massachusetts Youth Screening Instrument: identifying race and gender differential item functioning among juvenile offenders. *Educ Psychol Meas* 2006;**66**:502–21	5
25.	Chamberlain P, Saldana L, Brown CH, Leve LD. Implementation of multidimensional treatment foster care in California: a randomized control trial of an evidence-based practice. In Roberts-DeGennaro M, Foge SJ, editors. *Using Evidence to Inform Practice for Community and Organizational Change.* Chicago, IL: Lyceum Books; 2011. pp. 218–34	3B
26.	Chitsabesan P, Rothwell J, Kenning C, Law H, Carter L, Bailey S, *et al.* Six years on: a prospective cohort study of male juvenile offenders in secure care. *Eur J Child Adolesc Psychiatry* 2012;**21**:339–47	1
27.	Clarke G, Debar L, Lynch F, Powell J, Gale J, O'Connor E, *et al.* A randomized effectiveness trial of brief cognitive–behavioral therapy for depressed adolescents receiving antidepressant medication. *J Am Acad Child Adolesc Psychiatry* 2005;**44**:888–98	3B
28.	Cockett R. A short diagnostic self-rating scale in the pre-adult remand setting. *Br J Psychiatry* 1969;**115**:1141–50	5
29.	Cole DA. Validation of the Reasons for Living Inventory in general and delinquent adolescent samples. *J Abnorm Child Psychol* 1989;**17**:13–27	5
30.	Conners C, Kramer R, Rothschild G, Schwartz L, Stone A. Treatment of young delinquent boys with diphenylhydantoin sodium and methyphenidate. *Arch Gen Psychiatry* 1971;**24**:156–60	6B
31.	Connor D, Glatt S, Lopez I, Jackson D, Melloni R. Psychopharmacology and aggression. I: a meta-analysis of stimulant effects on overt/covert aggression-related behaviors. *J Am Acad Child Adolesc Psychiatry* 2002;**41**:253–61	6B
32.	Costello E, Copeland W, Cowell A, Keeler G. Service costs of caring for adolescents with mental illness in a rural community, 1993–2000. *Am J Psychiatry* 2007;**164**:36–42	3B

TABLE 37 List of excluded studies with reasons (*continued*)

No.	Study	Reason
33.	Crane DR, Hillin HH, Jakubowski SF. Costs of treating conduct disordered Medicaid youth with and without family therapy. *Am J Fam Ther* 2005;**33**:403–13	3B
34.	Crespi TD. *The Development of the Inventory of Adolescent Well-Being: A Follow-Up Study of the Effects of Psychiatric Hospitalization on Adolescents.* PhD thesis. Boston, MA: University of Massachusetts; 1986	3A
35.	Cromes GF. *The Effect of a Group Therapy Experience on Juvenile Offenders Relative to their Classification as Neurotic or Sociopathic.* PhD thesis. Lubbock, TX: Texas Tech University; 1972	6B
36.	Curtis NM, Ronan KR, Heiblum N, Crellin K. Dissemination and effectiveness of multisystemic treatment in New Zealand: a benchmarking study. *J Fam Psychol* 2009;**23**:119–29	3B
37.	Davidson W. *The Diversion of Juvenile Delinquents: an Examination of the Processes and Relative Efficacy of Child Advocacy and Behavioral Contracting.* Champaign, IL: University of Illinois; 1976	6B
38.	Davis BR. *Personality Characteristics Associated with Violence in Incarcerated Adolescent Males as Identified by Comparison of Offending Behavior and Millon Adolescent Clinical Inventory Scales.* PhD thesis. Chicago, IL: Adler School of Professional Psychology; 2003	1
39.	Deitche JH. *The Performance of Delinquent and Non-Delinquent Boys on the Tennessee Department of Mental Health Self-Concept Scale.* PhD thesis. Bloomington, IN: Indiana University; 1959	1
40.	Delisi M, Caudill JW, Trulson CR, Marquart JW, Vaughn MG, Beaver KM. Angry inmates are violent inmates: a Poisson regression approach to youthful offenders. *J Forensic Psychol Pract* 2010;**10**:419–39	5
41.	Dembo R, Anderson A. The problem oriented screening instrument for teenagers. In Grisso T, Vincent G, Seagrave D, editors. *Mental Health Screening and Assessment in Juvenile Justice.* New York, NY: Guilford Press; 2005. pp. 112–22	1
42.	Dembo R, Shemwell M, Guida J, Schmeidler J, Pacheco K, Seeberger W. A longitudinal study of the impact of a family empowerment intervention on juvenile offender psychosocial functioning: a first assessment. *J Child Adolesc Subst Abuse* 1998;**8**:15–54	1
43.	Dirks-Linhorst PA. *An Evaluation of a Family Court Diversion Program for Delinquent Youth with Chronic Mental Health Needs.* PhD thesis. Columbia, MO: University of Missouri; 2004	6B
44.	Dolin IH, Kelly DB, Beasley T. Chronic self-destructive behavior in normative and delinquent adolescents. *J Adolesc* 1992;**15**:57–66	3A
45.	Doreleijers T, Spaander M. The development and implementation of the BARO: a new device to detect psychopathology in minors with first police contacts. In Corrado R, Roesch R, Hart S, Gierowski J, editors. *Multi-Problem Violent Youth: a Foundation for Comparative Research on Needs, Interventions and Outcomes.* Amsterdam: IOS Press; 2002. pp. 232–40	5
46.	Doyle R, Mick E, Biederman J. Convergence between the Achenbach youth self-report and structured diagnostic interview diagnoses in ADHD and non-ADHD youth. *J Nerv Ment Dis* 2007;**195**:350–2	3A
47.	Duits N, Harkink J. [Usefulness of the DISC in juvenile forensic psychiatric diagnostics.] *Tijdschr Psychiatr* 2001;**43**:405–9	5
48.	Duncan RD, Kennedy WA, Patrick CJ. Four-factor model of recidivism in male juvenile offenders. *J Clin Child Psychol* 1995;**24**:250–7	1
49.	Earthrowl M, McCully R. Screening new inmates in a female prison. *J Forensic Psychiatry* 2002;**13**:428–39	3A
50.	Eckstein DG, Bailey WR. Investigating self-concept differences and utilizing group counseling with incarcerated and nonincarcerated juvenile delinquents. *J Humanics* 1977;**5**:106–11	2
51.	Ehlers A. Understanding and treating unwanted trauma memories in posttraumatic stress disorder. *Z Psychol* 2010;**218**:141–5	1
52.	Emshoff J, Blakely C. The diversion of delinquent youth: family focused intervention. *Child Youth Serv Rev* 1983;**5**:343–56	6B
53.	Esposito CL, Clum GA. Specificity of depression symptoms and suicidality in a juvenile delinquent population. *J Psychopathol Behav Assess* 1999;**21**:171–82	5

continued

TABLE 37 List of excluded studies with reasons (*continued*)

No.	Study	Reason
54.	Finnerty BM. *Utility of the Trauma Symptom Checklist for Children: Distinguishing between PTSD and Non-PTSD Youth in Residential Treatment.* PhD thesis. Chicago, IL: University of Chicago; 2002	3A
55.	Ford G, Andrews R, Booth A, Dibdin J, Hardingham S, Kelly T. Screening for learning disability in an adolescent forensic population. *J Forensic Psychiatry Psychol* 2008;**19**:371–81	6A
56.	Ford J, Steinberg K, Hawke J, Levine J, Zhang W. Randomized trial comparison of emotion regulation and relational psychotherapies for PTSD with girls involved in delinquency. *J Clin Child Adolesc Psychol* 2012;**41**:27–37	3B
57.	Foster E, Connor T. Public costs of better mental health services for children and adolescents. *Psychiatr Serv* 2005;**56**:50–5	1
58.	Foster E, Jones DE. The high costs of aggression: public expenditures resulting from conduct disorder. *Am J Public Health* 2005;**95**:1767–72	1
59.	Foster EM, Jensen PS, Schlander M, Pelham WE Jr, Hechtman L, Arnold LE, *et al.* Treatment for ADHD: is more complex treatment cost-effective for more complex cases? *Health Serv Res* 2007;**42**:165–82	3B
60.	Franklin G, Nottage W. Psychoanalytic treatment of severely disturbed juvenile delinquents in a therapy group. *Int J Group Psychother* 1969;**19**:165–75	1
61.	Friman P, Handwerk ML, Smith G, Larzelere R, Lucas C, Shaffer D. External validity of conduct and oppositional defiant disorder determined by the NIMH Diagnostic Interview Schedule for Children. *J Abnorm Child Psychol* 2000;**28**:277–86	3A
62.	Gavazzi SM, Lim JY, Yarcheck CM, Eyre EL. Brief report on predictive validity evidence of global risk indicators in the lives of court-involved youth. *Psychol Rep* 2003;**93**:1239–42	6A
63.	Gellman AR. *Guided Imagery: Exploring an Alternative Adjunctive Intervention with Traumatized Adolescents in Residential Foster Care.* PhD thesis. New York, NY: New York University; 2001	3B
64.	Gillikin CL. *Psychosocial Predictors of Juvenile Justice Involvement among Adolescent Female Offenders.* PhD thesis. Coral Gables, FL: University of Miami; 2010	5
65.	Glaser BA, Calhoun GB, Petrocelli JV, Bates JM, Owens-Hennick LA. Depression and somatic complaints among male juvenile offenders: differentiating somatizers from non-somatizers with the Millon Adolescent Clinical Inventory (MACI). *J Forensic Psychiatry Psychol* 2005;**16**:566–76	5
66.	Gleser G, Gottschalk L, Fox R, Lippert W. Immediate changes in affect with chlordiazepoxide. *Arch Gen Psychiatry* 1965;**13**:291–5	6B
67.	Glisson C, Hemmelgarn AL, Post JA. The Shortform Assessment for Children: an assessment and outcome measure for child welfare and juvenile justice. *Res Soc Work Pract* 2002;**12**:82–106	5
68.	Gottfredson DM, Snyder HN. *Mathematics of Risk Classification: Changing Data into Valid Instruments for Juvenile Court.* Washington, DC: Office of Juvenile Justice and Delinquency Prevention; 2005	6A
69.	Grant J. *A Problem Solving Intervention for Aggressive Adolescent Males: A Preliminary Investigation*. Syracuse, NY: Syracuse University; 1987	6B
70.	Gretton H, Clift R. The mental health needs of incarcerated youth in British Columbia, Canada. *Int J Law Psychiatry* 2011;**34**:109–15	1
71.	Griffith ML. *Experiential Therapy and Empathy in Oppositional Defiant Disorder Adolescents.* PhD thesis. Chicago, IL: Adler School of Professional Psychology; 2010	8
72.	Grisso T, Barnum R, Fletcher KE, Cauffman E, Peuschold D. Massachusetts Youth Screening Instrument for mental health needs of juvenile justice youths. *J Am Acad Child Adolesc Psychiatry* 2001;**40**:541–8	5
73.	Guerra N, Slaby R. Cognitive mediators of aggression in adolescent offenders: 2. Intervention. *Dev Psychol* 1990;**26**:269–77	6B
74.	Guevara JP, Mandell DS. Costs associated with attention deficit hyperactivity disorder: overview and future projections. *Expert Rev Pharmacoecon Outcomes Res* 2003;**3**:201–10	1
75.	Handwerk M, Friman P, Larzelere R. Comparing the DISC and the Youth Self-Report. *J Am Acad Child Adolesc Psychiatry* 2000;**39**:807	3B
76.	Harrington R, Bailey S. *Mental Health Needs and Effectiveness of Provision for Young Offenders in Custody and in the Community.* London: Youth Justice Board; 2005	1

TABLE 37 List of excluded studies with reasons (*continued*)

No.	Study	Reason
77.	Harrington TL. *The Clinician-Administered PTSD Scale for Children and Adolescents: A Validation Study.* PhD thesis. Tulsa, OK: University of Tulsa; 2009	5
78.	Harshbarger JL. *The Impact of Mental Health Dimensions on the Prediction of Juvenile Reentry Recidivism.* PhD thesis. Wichita, KS: Wichita State University; 2007	1
79.	Heckman CJ, Cropsey KL, Olds-Davis T. Posttraumatic stress disorder treatment in correctional settings: a brief review of the empirical literature and suggestions for future research. *Psychotherapy (Chic)* 2007;**44**:46–53	1
80.	Hemmelgarn AL, Glisson C, Sharp SR. The validity of the Shortform Assessment for Children (SAC). *Res Soc Work Pract* 2003;**13**:510–30	3A
81.	Henggeler S, Schoenwald S, Borduin C, Rowland M, Cunningham P. *Multisystemic Treatment of Antisocial Behaviour in Children and Adolescents.* New York, NY: Guilford; 1998	1
82.	Henggeler S, Rowland M, Randall J, Ward D, Pickrel S, Cunningham P, *et al.* Home-based multisystemic therapy as an alternative to the hospitalisation of youths in psychiatric crisis: clinical outcomes. *J Am Acad Child Adolesc Psychiatry* 1999;**38**:1331–9	3B
83.	Henggeler SW, Rodick J, Borduin CM, Hanson C, Watson S, Urey J. Multisystemic treatment of juvenile offenders: effects on adolescent behavior and family interaction. *Dev Psychol* 1986;**22**:132–41	8
84.	Henggeler SW, Melton GB, Smith LA. Family preservation using multisystemic therapy: an effective alternative to incarcerating serious juvenile offenders. *J Consult Clin Psychol* 1992;**60**:953–61	6B
85.	Henggeler SW, Melton GB, Brondino MJ, Scherer DG, Hanley JH. Multisystemic therapy with violent and chronic juvenile offenders and their families: the role of treatment fidelity in successful dissemination. *J Consult Clin Psychol* 1997;**65**:821–33	6B
86.	Hilyer JC, Wilson DG, Dillon C, Caro L, Jenkins C, Spencer WA, *et al.* Physical fitness training and counseling as treatment for youthful offenders. *J Couns Psychol* 1982;**29**:292–303	4
87.	Hussey DL, Drinkard AM, Falletta L, Flannery DJ. Understanding clinical complexity in delinquent youth: comorbidities, service utilization, cost, and outcomes. *J Psychoactive Drugs* 2008;**40**:85–95	1
88.	Hutt ML, Dates BG, Reid DM. The predicative ability of HABGT scales for a male delinquent population. *J Pers Assess* 1977;**41**:492–6	5
89.	Janus M-D. *Affective Empathy Training in the Treatment of Conduct Disordered Adolescents.* PhD thesis. Storrs, CT: University of Connecticut; 1993	8
90.	Jarden H. *A Comparison of Problem-Solving Interventions on the Functioning of Youth with Disruptive Behavior Disorder.* PhD thesis. Bethlehem, PA: Lehigh University; 1994	3B
91.	Jesness C. Comparative effectiveness of behavior modification and transactional analysis programs for delinquents. *J Consult Clin Psychol* 1975;**43**:758–79	6B
92.	Jesness CF. Comparative effectiveness of two institutional treatment programs for delinquents. *Child Care Q* 1971;**1**:119–30	6B
93.	Jewell J, Handwerk M, Almquist J, Lucas C. Comparing the validity of clinician-generated diagnosis of conduct disorder to the Diagnostic Interview Schedule for Children. *J Clin Child Adolesc Psychol.* 2004;**33**:536–46	3A
94.	Jones DE, Foster E. Service use patterns for adolescents with ADHD and comorbid conduct disorder. *J Behav Health Serv Res* 2009;**36**:436–49	3B
95.	Jurjevich RRM. *No Water in My Cup: Experiences and a Controlled Study of Psychotherapy of Delinquent Girls.* New York: Libra; 1968	8
96.	Kazdin A. *Psychotherapy for Children and Adolescents: Directions for Research and Practice.* New York: Oxford University Press; 2000	1
97.	Klarreich S. A study comparing two treatment approaches with adolescent probationers. *Correct Soc Psychiatry J Behav Technol Methods Ther* 1981;**11**:101–14	6B
98.	Klietz SJ, Borduin CM, Schaeffer CM. Cost–benefit analysis of multisystemic therapy with serious and violent juvenile offenders. *J Fam Psychol* 2010;**24**:657–66	6B

continued

TABLE 37 List of excluded studies with reasons (*continued*)

No.	Study	Reason
99.	Knapp M. Economic evaluations and interventions for children and adolescents with mental health problems. *J Child Psychol Psychiatry* 1997;**38**:3–25	1
100.	Knapp M. Hidden costs of mental illness. *Br J Psychiatry* 2003;**183**:477–8	1
101.	Knight J, Goodman E, Pulerwitz T, DuRant RH. Reliability of the Problem Oriented Screening Instrument for Teenagers (POSIT) in adolescent medical practice. *J Adolesc Health* 2001;**29**:125–30	3A
102.	Knott JM. *Self-Efficacy and Motivation to Change among Chronic Youth Offenders: An Exploratory Examination of the Efficacy of an Experiential Learning Motivation Enhancement Intervention.* PhD thesis. Eugene, OR: University of Oregon; 2004	6B
103.	Kohler D, Hinrichs G, Otto T, Huchzermeier C. On the psychological distress of young prison inmates (measured by the SCL-90-R). *Recht Psychiatrie* 2004;**22**:138–42	1
104.	Koles MR. *Validation and Use of the Child Behavior Checklist with Adolescent Person and Property Offenders Referred to a Mental Health Unit.* PhD thesis. Salt Lake City, UT: University of Utah; 1989	5
105.	Krakow B, Sandoval D, Schrader R, Keuhne B, McBride L, Yau C, *et al.* Treatment of chronic nightmares in adjudicated adolescent girls in a residential facility. *J Adolesc Health* 2001;**29**:94–100	8
106.	Kroll L. Needs assessment in adolescent offenders. In Bailey S, Dolan M, editors. *Adolescent Forensic Psychiatry.* London: Hodder Arnold; 2004. pp. 14–26	1
107.	Kroll L, Woodham A, Rothwell J, Bailey S, Tobias C, Harrington R, *et al.* Reliability of the Salford Needs Assessment Schedule for Adolescents. *Psychol Med* 1999;**29**:891–902	3A
108.	Kroll L, Harrington R, Bailey S. Needs assessment of children and adolescents. *Child Psychol Psychiatry Rev* 2000;**5**:81–8	1
109.	Lee R, Haynes NM. Counseling juvenile offenders: an experimental evaluation of Project CREST. *Comm Ment Health J* 1978;**14**:267–71	6B
110.	Leeman LW, Gibbs JC, Fuller D. Evaluation of a multi-component group treatment program for juvenile delinquents. *Aggres Behav* 1993;**19**:281–92	6B
111.	Leschied A, Cunningham A. *Seeking Effective Interventions for Serious Young Offenders: Interim Results of Four-Year Randomized Study of Multisystemic Therapy in Ontario, Canada.* Ontario: Center for Children and Families in the Justice System; 2002	6B
112.	Leve LD, Chamberlain P. Girls in the juvenile justice system: risk factors and clinical implications. In Pepler DJ, Madsen KC, Webster C, Levene K, editors. *The Development and Treatment of Girlhood Aggression.* Mahwah, NJ: Lawrence Erlbaum Associates Publishers; 2005. pp. 191–215	6B
113.	Littell JH, Campbell M, Green S, Toews B. Multisystemic therapy for social, emotional, and behavioral problems in youth aged 10–17. *Cochrane Database Syst Rev* 2005;**4**:CD004797	1
114.	Lodewijks HP, Doreleijers TA, de Ruiter C. SAVRY risk assessment in violent Dutch adolescents: relation to sentencing and recidivism. *Crim Justice Behav* 2008;**35**:696–709	6A
115.	Lodewijks HPB, Doreleijers TAH, de Ruiter C, Borum R. Predictive validity of the Structured Assessment of Violence Risk in Youth (SAVRY) during residential treatment. *Int J Law Psychiatry* 2008;**31**:263–71	6A
116.	Loney BR, Counts CA. Scales for Assessing Attention-Deficit/Hyperactivity Disorder. In Grisso T, Vincent G, Seagrave D, editors. *Mental Health Screening and Assessment in Juvenile Justice.* New York: Guilford Press; 2005. pp. 166–84	1
117.	Lucas C, Zhang H, Fisher P, Shaffer D, Regier D, Narrow W, *et al.* The Disc Predictive Scales: efficiently screening for diagnoses. *J Am Acad Child Adolesc Psychiatry* 2001;**40**:443–9	3A
118.	MacMahon JR, Gross RT. Physical and psychological effects of aerobic exercise in delinquent adolescent males. *Am J Dis Child* 1988;**142**:1361–6	8
119.	Maggiolini A, Ciceri A, Pisa C, Belli S. Mental health problems in young offenders. *Infanzia Adolesc* 2009;**8**:139–50	1
120.	Martsch MD. *A Comparison of Two Cognitive–Behavioral Group Treatments for Adolescent Aggression: High-Process Versus Low-Process.* PhD thesis. Madison, WI: University of Wisconsin-Madison; 2000	3B
121.	McConville DW. *The Millon Adolescent Clinical Inventory in the Assessment of Juvenile Offenders.* PhD thesis. Charlottesville, VA: University of Virginia; 2004	5

TABLE 37 List of excluded studies with reasons (*continued*)

No.	Study	Reason
122.	McLaney M, Del Boca F, Babor T. A validation of the problem oriented screening instrument for teenagers (POSIT). *J Ment Health* 1994;**363**:363–76	3A
123.	Meier R, Dittrich A, Andreae A. Psychiatric prognosis in criminal cases involving older teenagers and young adults. *Schweizer Archiv Neurol Psychiatrie* 2004;**155**:264–72	3B
124.	Mendonsa AD. *Sacramento Assessment Center: Using Comprehensive Multidimensional Assessments in Increasing Juvenile Offender Placement Success and Reducing Recidivism.* PhD thesis. San Diego, CA: Alliant International University; 2008	8
125.	Miller JB. *The Effects of a Cognitive–Behavioral Group Intervention on Depressive Symptoms in an Incarcerated Adolescent Delinquent Population (Juvenile Delinquents).* PhD thesis. Berkeley, CA: Wright Institute; 1999	8
126.	Mohino Justes S, Ortega-Monasterio L, Planchat Teruel LM, Cuquerella Fuentes A, Talon Navarro T, Macho Vives LJ. Discriminating deliberate self-harm (DSH) in young prison inmates through personality disorder. *J Forensic Sci* 2004;**49**:137–40	3A
127.	Moore KJ, Sprengelmeyer PG, Chamberlain P. Community-based treatment for adjudicated delinquents: the Oregon Social Learning Center's 'Monitor' Multidimensional Treatment Foster Care Program. *Resid Treat Child Youth* 2001;**18**:87–97	1
128.	Moore RH. Construct validity of the MacAndrew scale: secondary psychopathic and dysthymic-neurotic character orientations among adolescent male misdemeanor offenders. *J Stud Alcohol* 1985;**46**:128–31	5
129.	Morris J. *The Effectiveness of Anger-Control Training with Institutionalized Juvenile Offenders: the 'Keep Cool' Program.* Richmond, VA: Virginia Commonwealth University; 1981	6B
130.	Mulvey E, Arthur M, Reppucci N. The prevention and treatment of juvenile delinquency: a review of the research. *Clin Psychol Rev* 1993;**13**:133–67	1
131.	Murrie DC, Cornell DG. Psychopathy screening of incarcerated juveniles: a comparison of measures. *Psychol Assess* 2002;**14**:390–6	6A
132.	Myers WC, Burton PR, Sanders PD, Donat KM, Cheney J, Fitzpatrick TM, *et al.* Project Back-on-Track at 1 year: a delinquency treatment program for early-career juvenile offenders. *J Am Acad Child Adolesc Psychiatry* 2000;**39**:1127–34	8
133.	Nakaya N, Kumano H, Minoda K, Koguchi T, Tanouchi K, Kanazawa M, *et al.* Preliminary study: psychological effects of muscle relaxation on juvenile delinquents. *Int J Behav Med* 2004;**11**:176–80	6B
134.	National Center for Mental Health and Juvenile Justice. *Screening & Assessing Mental Health & Substance Use Disorders among Youth in the Juvenile Justice System: A Resource Guide for Practitioners*. Delmar, NY: National Center for Mental Health and Juvenile Justice; 2004	1
135.	Newman E, Kaloupek D. Posttraumatic stress disorder among criminally involved youth. *Arch Gen Psychiatry* 2003;**60**:849	1
136.	Nordness PD, Grummert M, Banks D, Schindler ML, Moss MM, Gallagher K, *et al.* Screening the mental health needs of youths in juvenile detention. *Juv Fam Court J* 2002;**53**:43–50	5
137.	Ogden T, Halliday-Boykins C. Multisystemic treatment of antisocial adolescents in Norway: replication of clinical outcomes outside of the US. *Child Adolesc Ment Health* 2004;**9**:77–83	3B
138.	Ogden T, Amlund Hagen K. Multisystemic treatment of serious behaviour problems in youth: sustainability of effectiveness two years after intake. *Child Adolesc Ment Health* 2006;**11**:142–9	3B
139.	Ogden T, Hagen KA, Anderson O. Sustainability of the effectiveness of a programme of multisystemic treatment (MST) across participant groups in the second year of operation. *J Child Serv* 2007;**2**:4–14	3B
140.	Okazaki M. *Identification of Self-Destructiveness among Incarcerated Juvenile Populations.* PhD thesis. San Diego, CA: Alliant International University; 1996	1
141.	Olsson TM. Intervening in youth problem behavior in Sweden: a pragmatic cost analysis of MST from a randomized trial with conduct disordered youth. *Int J Soc Welfare* 2010;**19**:194–205	3B
142.	O'Malley EM. *Client Characteristics that are Associated with Positive Treatment Outcomes for a Delinquent Population that Participates in a community Based Treatment: an Archival Review.* PhD thesis. Chicago, IL: Chicago School of Professional Psychology; 2008	1

continued

TABLE 37 List of excluded studies with reasons (*continued*)

No.	Study	Reason
143.	Ovaert LB, Cashel ML, Sewell KW. Structured group therapy for posttraumatic stress disorder in incarcerated male juveniles. *Am J Orthopsychiatry* 2003;**73**:294–301	8
144.	Partridge CR. *Concurrent Validity of Parent Reports Regarding the Family/Parenting Dimension of a Global Risk Assessment Device for Court-Involved Adolescents and Their Families.* PhD thesis. Columbus, OH: Ohio State University; 2008	6A
145.	Pelham WE, Foster E, Robb JA. The economic impact of attention-deficit/hyperactivity disorder in children and adolescents. *J Pediatr Psychol* 2007;**32**:711–27	1
146.	Penner E, Roesch R, Viljoen J. Young offenders in custody: an international comparison of mental health services. *Int J Forensic Ment Health* 2011;**10**:215–32	1
147.	Perry AE, Gilbody S. Detecting and predicting self-harm behaviour in prisoners: a prospective psychometric analysis of three instruments. *Soc Psychiatry Psychiatr Epidemiol* 2009;**44**:853–61	3A
148.	Perry AE, Olason DT. A new psychometric instrument assessing vulnerability to risk of suicide and self-harm behaviour in offenders: Suicide Concerns for Offenders In Prison Environment (SCOPE). *Int J Offend Ther Comp Criminol* 2009;**53**:385–400	3A
149.	Pineda DA, Kamphaus RW, Restrepo MA, Puerta IC, Arango CP, Lopera FJ, *et al.* Screening for conduct disorder in an adolescent male sample from Colombia. *Transcult Psychiatry* 2006;**43**:362–82	3A
150.	Poland AL, Diaz S. *Examination of the Usefulness of a Suicide Checklist in a Predominantly Hispanic Juvenile Detention Population.* American Society of Criminology 62nd Annual Meeting, San Francisco, CA, 17–20 November 2010	2
151.	Pomeroy EC, Green DL, Kiam R. Female juvenile offenders incarcerated as adults: a psychoeducational group intervention. *J Soc Work* 2001;**1**:101–15	8
152.	Rand MR. *The Effectiveness of the Bender Gestalt Test in Differentiating Adjustment Disorder, Attention Deficit Disorder, and Conduct Disorder in Children.* PhD thesis. San Francisco, CA: Saybrook Institute; 1987	3A
153.	Reardon JP, Tosi DJ. The effects of rational stage directed imagery on self-concept and reduction of psychological stress in adolescent delinquent females. *J Clin Psychol* 1977;**33**:1084–92	6B
154.	Regan TP. *An Analysis and Validation of Scale-8 on the MMPI-168 with Juvenile Delinquents.* PhD thesis. Detroit, MI: Wayne State University; 1992	5
155.	Rennie C, Dolan M. Predictive validity of the youth level of service/case management inventory in custody sample in England. *J Forensic Psychiatry Psychol* 2010;**21**:407–25	6A
156.	Reppucci ND, Redding RE. *Screening Instruments for Mental Illness in Juvenile Offenders: The MAYSI and the BSI (Juvenile Justice Fact Sheet).* Charlottesville, VA: University of Virginia, Institute of Law, Psychiatry, and Public Policy; 2000	5
157.	Rogers KL. *Posttraumatic Stress Disorder in a Sample of Conduct Disordered Youth.* PhD thesis. Burnaby, BC: Simon Fraser University; 1996	1
158.	Rohde P, Seeley JR, Kaufman NK, Clarke GN, Stice E. Predicting time to recovery among depressed adolescents treated in two psychosocial group interventions. *J Consult Clin Psychol* 2006;**74**:80–8	1
159.	Saxena K, Silverman MA, Chang K, Khanzode L, Steiner H. Baseline predictors of response to divalproex in conduct disorder. *J Clin Psychiatry* 2005;**66**:1541–8	6B
160.	Schlicter K, Horan J. Effects of stress inoculation on the anger and aggression management skills of institutionalized juvenile delinquents. *Cog Ther Res* 1981;**5**:359–65	6B
161.	Schmidt F, Hoge RD, Gomes L. Reliability and validity analyses of the youth level of service/case management inventory. *Crim Justice Behav* 2005;**32**:329–44	6A
162.	Scott S, Knapp M, Henderson J, Maughan B. Financial cost of social exclusion: follow up study of anti-social children into adulthood. *BMJ* 2001;**323**:191–4	1
163.	Sexton T, Turner CW. The effectiveness of functional family therapy for youth with behavioral problems in a community practice setting. *J Fam Psychol* 2010;**24**:339–48	6B
164.	Shelton D. Patterns of treatment services and costs for young offenders with mental disorders. *J Child Adolesc Psychiatric Nurs* 2005;**18**:103–12	1

TABLE 37 List of excluded studies with reasons (*continued*)

No.	Study	Reason
165.	Smith CA. *The Effects of a Cognitive–Behavioral Prevention Program on Social Skills and Interpersonal Guilt: a Study of Delinquent Adolescents.* PhD thesis. Berkeley, CA: Wright Institute; 1998	6B
166.	Smith DJ, Michael R, David JS. Youth crime and conduct disorders: trends, patterns and causal explanations. In Rutter M, Smith DJ, editors. *Psychosocial Disorders in Young People: Time Trends & Their Causes.* Chichester: Wiley; 1995. pp. 389–489	1
167.	Snyder J, White M. The use of cognitive self-instruction in the treatment of behaviorally disturbed adolescents. *Behav Ther* 1979;**10**:1409–16	3B
168.	Stathis S, Litchfield B, Letters P, Doolan I, Martin G. A comparative assessment of suicide risk for young people in youth detention. *Arch Suicide Res* 2008;**12**:62–6	5
169.	Stein LAR, Lebeau-Craven R, Martin R, Colby SM, Barnett NP, Golembeske C Jr, *et al.* Use of the adolescent SASSI in a juvenile correctional setting. *Assessment* 2005;**12**:384–94	6A
170.	Steiner H, Garcia I, Matthews Z. Posttraumatic stress disorder in incarcerated juvenile delinquents. *J Am Acad Child Adolesc Psychiatry* 1997;**36**:357–65	1
171.	Steiner H, Petersen ML, Saxena K, Ford S, Matthews Z. Divalproex sodium for the treatment of conduct disorder: a randomized controlled clinical trial. *J Clin Psychiatry* 2003;**64**:1183–91	6B
172.	Stewart C, Rapp-Paglicci L, Rowe W. Evaluating the efficacy of the prodigy prevention program across urban and rural locales. *Child Adolesc Soc Work J* 2009;**26**:65–75	8
173.	Sundell K, Hansson K, Lofholm CA, Olsson T, Gustle L-H, Kadesjo C. The transportability of multisystemic therapy to Sweden: short-term results from a randomized trial of conduct-disordered youths. *J Fam Psychol* 2008;**22**:550–60	3B
174.	Swanson J, Sergeant J, Taylor E, Sonuga-Barke E, Jensen P, Cantwell D. Attention deficit hyperactivity disorder and hyperkinetic disorder. *Lancet* 1998;**351**:429–33	1
175.	Taylor A. An evaluation of group psychotherapy in a girls' borstal. *Int J Group Psychother* 1967;**17**:168–77	2
176.	Tellier JE. *Anger and Depression among Incarcerated Juvenile Delinquents: A Pilot Intervention.* PhD thesis. Berkeley, CA: Wright Institute; 1999	8
177.	Tennyson DH. Juvenile correctional system health care costs: a five-year comparison. *J Correct Health Care* 2003;**10**:257–71	1
178.	Tennyson DH. Predicting medication costs and usage: expenditures in a juvenile detention facility. *J Correct Health Care* 2009;**15**:98–104	1
179.	Timmons-Mitchell J, Bender MB, Kishna MA, Mitchell CC. An independent effectiveness trial of multisystemic therapy with juvenile justice youth. *J Clin Child Adolesc Psychol* 2006;**35**:227–36	6B
180.	Townsend E, Walker D-M, Sargeant S, Stocker O, Vostanis P, Sithole J, *et al.* Interventions for mood and anxiety disorders, and self harm in young offenders. *Cochrane Database Syst Rev* 2008;**2**:CD007195	1
181.	Townsend E, Walker D-M, Sargeant S, Vostanis P, Hawton K, Stocker O, *et al.* Systematic review and meta-analysis of interventions relevant for young offenders with mood disorders, anxiety disorders, or self-harm. *J Adolesc* 2010;**33**:9–20	1
182.	Tranah T, Hill A. Assessment of delinquent adolescents using Achenbach's Teacher's Report Form. *Pers Individ Dif* 2000;**29**:109–17	3A
183.	Tuton FL, Hood JW. Psychotherapy in criminal rehabilitation: individual vs. group psychotherapy in relation to post-therapy recidivism. *J Comm Correct Centers* 1973;**1**:11–18	2
184.	van Lier PAC, Verhulst FC, Crijnen AAM. Screening for disruptive behavior syndromes in children: the application of latent class analyses and implications for prevention programs. *J Consult Clin Psychol* 2003;**71**:353–63	3A
185.	Vaughan PJ. Secure care and treatment needs of mentally disordered adolescents. *Br J Forensic Pract* 2004;**6**:14–20	1
186.	Velasquez JS, Lyle CG. Day versus residential treatment for juvenile offenders: the impact of program evaluation. *Child Welfare* 1985;**54**:145–56	6B

continued

TABLE 37 List of excluded studies with reasons (*continued*)

No.	Study	Reason
187.	Vinick B. *The Effects of Assertiveness Training on Aggression and Self-Concept in Conduct Disordered Adolescents.* Memphis, TN: Memphis State University; 1983	3B
188.	von Sydow K, Beher S, Schweitzer-Rothers J, Retzlaff R. Systemic family therapy with children and adolescents as index patients. A meta-content analysis of 47 randomized controlled outcome studies. *Psychotherapeut* 2006;**51**:107–43	1
189.	Washington State Institute for Public Policy. *Washington State's Family Integrated Transitions Program for Juvenile Offenders: Outcome Evaluation & Benefit–Cost Analysis*. Olympia, WA: Washington State Institute for Public Policy; 2004	6B
190.	Wasserman GA, Ko SJ, McReynolds LS. Assessing the mental health status of youth in juvenile justice settings. *Juvenile Justice Bulletin*, August 2004. URL: www.ncjrs.gov/pdffiles1/ojjdp/202713.pdf (accessed 21 September 2014)	2
191.	Wasserman GA, Vilhauer JS, McReynolds L, Shoai R, Jonh R. Mental health screening in the juvenile justice system: a comparison between the VOICE-DISC-IV and the MAYSI-2. *J Juv Just Serv* 2004;**19**:7–17	7
192.	Wasserman GA, McReynolds LS, Fisher P, Lucas CP. Diagnostic Interview Schedule for Children: Present State Voice Version. In Grisso T, Vincent G, Seagrave D, editors. *Mental Health Screening and Assessment in Juvenile Justice.* New York: Guilford Press; 2005. pp. 224–39	1
193.	Wells C. The Treatment of Severe Antisocial Behaviour in Young People. In Baruch G, editor. *Community-Based Psychotherapy with Young People: Evidence and Innovation in Practice.* New York: Brunner-Routledge; 2001. pp. 128–41	6B
194.	White SF. *Examining the Influence of Callous-Unemotional Traits on Outcomes in an Evidence-Based Treatment Program for Delinquent Adolescents.* PhD thesis. New Orleans, LA: University of New Orleans; 2011	8
195.	Williams V, Grisso T, Valentine M, Remsburg N. Mental health screening: Pennsylvania's experience in juvenile detention. *Correct Today* 2008;**70**:24–7	5
196.	Wolff JC, Greene RW, Ollendick TH. Differential responses of children with varying degrees of reactive and proactive aggression to two forms of psychosocial treatment. *Child Fam Behav Ther* 2008;**30**:37–50	3B
197.	Woolfenden S, Williams Katrina J, Peat J. Family and parenting interventions in children and adolescents with conduct disorder and delinquency aged 10-17. *Cochrane Database Syst Rev* 2001;**2**:CD003015	1
198.	Young S, Gudjonsson G, Misch P, Collins P, Carter P, Redfern J, *et al.* Prevalence of ADHD symptoms among youth in a secure facility: the consistency and accuracy of self- and informant-report ratings. *J Forensic Psychiatry Psychol* 2010;**21**:238–46	1
199.	Youth Justice Board. *ASSET.* London: Youth Justice Board; nd	6A
200.	Youth Justice Board. *Screening for Mental Disorder in the Youth Justice System: Supporting Notes.* London: Youth Justice Board; 2003	1

Appendix 5 Quality Assessment of Diagnostic Accuracy Studies – version 2 field guide

Background

This guide deals with the use of QUADAS-2 in the young offender review. Whiting *et al.*[33] state that 'The QUADAS-2 tool must be tailored to each review by adding or omitting signalling questions and developing review-specific guidance on how to assess each signalling question and use this information to judge the risk of bias.' This guide aims to tailor QUADAS-2 to the young offender review.

Review question

The review question can be broken down as follows:

- patients: young people (aged 10–21 years) who have offended and who are in contact with the criminal justice system
- index test and target condition: screening measures designed to identify one or more of the following: depression, anxiety problems (including PTSD), risk of self-harm/suicide, schizophrenia, psychosis, ADHD and conduct disorders, autism spectrum disorders and bipolar disorder
- reference tests: for studies reporting diagnostic accuracy, a standardised diagnostic interview conducted to internationally recognised criteria (e.g. ICD or DSM).

The following additional information is required for QUADAS-2:

- setting: any part of the criminal justice system
- intended use of the index test: to identify a target condition or a mental health need
- patient presentation: patients can be tested at any point during their contact with the criminal justice system
- previous testing: patients may or may not have been tested previously.

The review question has been defined and it is now possible to review how each of the four QUADAS-2 domains can be applied to the young offender review

Domain 1: patient selection

Risk of bias: could the selection of patients have introduced bias?

Signalling question 1: was a consecutive or random sample of patients enrolled?

This question is relevant to the review. Consider whether the patients approached to take part represented a consecutive or a random sample. If this is not the case the question should be rated as 'no'. Non-random sampling and patients who refuse to participate or who drop out before enrolment will affect the randomness of the sample.

Signalling question 2: was a case–control design avoided?

This question is relevant to the review. Studies that enrol participants known to have the mental health problem in question and that enrol a control group known not to have the problem may exaggerate diagnostic accuracy.

Signalling question 3: did the study avoid inappropriate exclusions?

This question is relevant to the review and relates to potential participants excluded by the investigators. Consider the reasons given for exclusions and how many potential participants have been excluded. Although any exclusion theoretically introduces the potential for bias, the decision may affect very few potential participants.

How to rate: if any of the three questions is rated as 'no' there is a high risk of bias. If all three questions are rated as 'yes' there is a low risk of bias. If any of the questions are reported as 'unclear', then there is an unclear risk of bias and a judgement should be made on whether or not there is enough information to make a decision about the risk of bias.

Applicability: is there concern that the included patients do not match the review question?

We anticipate that many of the diagnostic studies will relate to only a proportion of the patients in the review question as we have developed very broad inclusion criteria. As long as the included patients match some of the inclusion criteria, this should be scored as *low*, although the discrepancy should be highlighted in the review.

Domain 2: index test(s)

Risk of bias: could the conduct or interpretation of the index test have introduced bias?

Signalling question 1: were the index tests interpreted without knowledge of the results of the reference standard?

This question is relevant to the review. We anticipate that the answer to this question will often be 'yes' as most of the assessments will be self-reported and therefore not prone to assessor bias. However, consider whether or not the questions need to be read out to a substantial proportion of the participants because of literacy difficulties.

Signalling question 2: if a threshold was used, was it prespecified?

This question is relevant to the review. Cut-off points for a rating scale should be specified a priori and this must be clearly stated in the paper. If this is not the case, the answer is 'no'.

How to rate: if either of the two questions is rated as 'no' there is a *high* risk of bias. If both questions are rated as 'yes' there is a low risk of bias. If any of the signalling questions is reported as 'unclear' there is an *unclear* risk of bias and a judgement should be made on whether there is enough information to make a decision about the risk of bias.

Applicability: is there concern that the index test, its conduct or its interpretation differ from the review question?

The index test should relate to at least one of the conditions specified or be an assessment of a mental health need.

Domain 3: reference standard

Risk of bias: could the reference standard, its conduct or its interpretation have introduced bias?

Signalling question 1: is the reference standard likely to correctly classify the target condition?

This question is relevant to the review. The reference standard should be a diagnostic interview and should be carried out by a qualified person. The person conducting the interview must have been appropriately trained, have had their performance satisfactorily benchmarked or have rated well on inter-rater reliability tests. If none of these conditions has been met, the answer is 'no'.

Signalling question 2: were the reference standard results interpreted without knowledge of the results of the index test?

This question is relevant to the review. We anticipate that the answer to this question will often be 'yes' as most of the assessments will be self-reported and therefore not prone to assessor bias. However if a researcher or clinician undertakes the diagnostic interview, it must be clear that he or she was blind to the results of the index test.

How to rate: if either of the questions is rated as 'no' there is a *high* risk of bias. If both of the questions are rated as 'yes' there is a *low* risk of bias. If any of the signalling questions is reported as 'unclear' there is an *unclear* risk of bias and a judgement should be made on whether or not there is enough information to make a decision about the risk of bias.

Applicability: are there concerns that the target condition as defined by the reference standard does not match the question?

The reference test should relate to at least one of the conditions specified or validate the presence or absence of a mental health need.

Domain 4: flow and timing

Risk of bias: could the patient flow have introduced bias?

Signalling question 1: was there an appropriate interval between the index test and the reference standard?

This question is relevant to the review. In the case of a diagnostic assessment, the index test and reference test must be conducted within 2 weeks of each other for this item to be rated 'yes'.

Signalling question 2: did all patients receive a reference standard?

Signalling question 3: did all patients receive the same reference standard?

These questions are relevant to the review and are self-explanatory. Consider if any reasons given for not giving all patients the same reference test are reasonable and whether or not the differences could have introduced bias.

Signalling question 4: were all patients included in the analysis?

This question is relevant to the review. There must be complete data for at least 90% of the patients enrolled in the study for this question to be rated 'yes'. If there is < 90% or evidence of a systematic difference between those with complete follow-up data and those without, this question should be rated as 'no'.

How to rate: if any of the four questions is rated as 'no' there is a high risk of bias. If all four questions are rated as 'yes' there is a *low* risk of bias. If any of the signalling questions is reported as 'unclear' there is an *unclear* risk of bias and a judgement should be made on whether or not there is enough information to make a decision about the risk of bias.

Appendix 6 Further details of the economic analysis

TABLE 38 Data extraction from Revicki and Wood[81] to estimate DFDs and associated days at full health to indicate incremental QALYs

Weeks	Proportion of group depressed (D_t)[a]		Proportion of time depressed within period (DepAv)[b]		Estimated days depressed (DDs)[c]		DFDs[d]		Utility state associated with being not depressed (U_Dep = 0) and mildly depressed (U_Dep = 1)[e]		Number of days with full quality of life[f]	
Time(t)	CWD-A	LS	CWD-A	LS	CWD-A	LS	CWD-A	LS	U_Dep = 0	U_Dep = 1	CWD-A	LS
0	1	1	0.8	0.905	22.4	25.34	5.6	2.66	0.85	0.685	20.104	19.6189
4	0.6	0.81	0.59	0.77	16.52	21.56	11.48	6.44	0.85	0.685	21.0742	20.2426
8	0.58	0.73	0.52	0.665	14.56	18.62	13.44	9.38	0.85	0.685	21.3976	20.7277
12	0.46	0.6	0.45	0.55	12.6	15.4	15.4	12.6	0.85	0.685	21.721	21.259
16	0.44	0.5	0.415	0.48	11.62	13.44	16.38	14.56	0.85	0.685	21.8827	21.5824
20	0.39	0.46	0.365	0.43	10.22	12.04	17.78	15.96	0.85	0.685	22.1137	21.8134
24	0.34	0.4	0.34	0.39	9.52	10.92	18.48	17.08	0.85	0.685	22.2292	21.9982
28	0.34	0.38	0.33	0.36	9.24	10.08	18.76	17.92	0.85	0.685	22.2754	22.1368
32	0.32	0.34	0.305	0.315	8.54	8.82	19.46	19.18	0.85	0.685	22.3909	22.3447
36	0.29	0.29	0.27	0.28	7.56	7.84	20.44	20.16	0.85	0.685	22.5526	22.5064
40	0.25	0.27	0.235	0.27	6.58	7.56	21.42	20.44	0.85	0.685	22.7143	22.5526
44	0.22	0.27	0.22	0.27	6.16	7.56	21.84	20.44	0.85	0.685	22.7836	22.5526
48	0.22	0.27	0.21	0.27	5.88	7.56	22.12	20.44	0.85	0.685	22.8298	22.5526
52	0.2	0.27	0.2	0.27	5.6	7.56	22.4	20.44	0.85	0.685	22.876	22.5526
56	0.2	0.27	0.2	0.25	5.6	7	22.4	21	0.85	0.685	22.876	22.645
60	0.2	0.23	0.175	0.23	4.9	6.44	23.1	21.56	0.85	0.685	22.9915	22.7374
64	0.15	0.23	0.15	0.23	4.2	6.44	23.8	21.56	0.85	0.685	23.107	22.7374

CWD-A, group treated with CBT; LS, control group.
a Extracted from Rohde *et al.*,[64] *Figure 2* (Kaplan–Meier product limit survival analysis).
b DepAv = (Dep_t – Dep_t + 1)/2.
c DDs = interval length (days) × DepAv.
d DFDs = interval length (days) – DDs.
e Revicki and Wood.[81]
f (DFDs × U_Dep = 0) + (DDs × U_Dep = 1).

TABLE 39 Costs associated with reoffending by crime type and cost per crime utilised to calculate the average cost of crime

Types of reoffences	Cost per crime[88] (£)		Reoffences (2013),[84] *n* (%)	Total cost of reoffences (£)
	Mean	95% CI		
All vehicle crime	1898	1.8 to 2069		
Taking and driving away	4800	3700 to 5500	1299 (4.05)	6,236,615
Theft from vehicle	580	570 to 620	542 (1.69)	314,466
Other motoring offences (all vehicle crime cost taken)	1791	1791 to 1952	3351 (10.44)	6,001,838
Drink driving (all vehicle crime cost taken)	1791	1791 to 1952	183 (0.57)	326,881
Non-vehicle theft (cost of handling)	725	704 to 725	791 (2.46)	573,335
Violent crime	38,393	29,861 to 46,924		
Serious wounding	130,000	100,000 to 160,000	293 (0.91)	38,050,006
Non-serious wounding	2000	1700 to 2200	8747 (27.24)	17,494,578
Sexual offences (number combines child and non-child offences)	40,526	9172 to 319,939	171 (0.53)	6,915,587
Sum of robberies to individuals and to premises	10,043	6886 to 50,466	1231 (3.83)	12,365,307
Theft or criminal damage				
Burglary in a dwelling	4906	4692 to 5332	1684 (5.24)	8,260,233
Burglary not in a dwelling	5759	5546 to 5759	1318 (4.10)	7,588,255
Theft from a shop	213	107 to 235	6882 (21.43)	1,467,892
Criminal damage against commercial/public sector property	1898	640 to 1898	5618 (17.50)	10,665,438

EME
HS&DR
HTA
PGfAR
PHR

Part of the NIHR Journals Library
www.journalslibrary.nihr.ac.uk

Published by the NIHR Journals Library