Volume One

Using Patient Profiling Tools to Predict and Enhance Therapy Outcomes

Katherine Garzonis

D.Clin.Psy. thesis (Volume 1), 2019

University College London

UCL DOCTORATE IN CLINICAL PSYCHOLOGY

THESIS DECLARATION FORM

I confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signature:

Name: Katherine Garzonis

Date:

ACKNOWLEDGEMENTS

Huge thanks to my supervisors, Professor Stephen Pilling and Doctor Amanda C de C Williams, and to Doctor Robert Saunders for their guidance and taking the time to help me through this. Acknowledgements are not enough.

Much gratitude and love to Stavros for his understanding and letting me leave piles of paper all over the house. Appreciation to Alex for not eating those.

Awe to my family for not only offering to read this, but actually doing so and giving some incredible feedback.

And a smile to little Leo, my actual thesis baby.

OVERVIEW

Algorithm-based decisions and personalised treatment are two major contemporary healthcare trends. This thesis investigates their utility and potential impact.

Part One is a literature review on the effectiveness of predictive decision support algorithms to improve mental health outcomes. It examines thirty papers to indicate their efficacy in practice and risks to uptake. Overall these systems are effective, but have a number of practical and psychological barriers to overcome to be implemented successfully. The review summarises eight hypotheses for effectiveness, which act as guidelines for designing future decision support systems.

Part Two examines a previously developed decision support algorithm and its ability to influence mental health recovery and improvement rates through individualised therapy allocation. Several ways of modelling the original algorithm are developed and compared on their ability to predict clinical outcomes, and then used to investigate historical trends in recovery rates at a particular service. Over time, service clinicians appear to naturally allocate more appropriate therapeutic intensities. Allocation as usual was compared with the decision support algorithm for clinical utility. The algorithm did not improve clinical outcomes but was more cost-effective.

Part Three is a critical appraisal and reflection on the research process in the context of wider technological and epistemological trends. It discusses the past role of people in research and how they may be involved in future scientific discovery given rapid advances in automating research procedures. It then examines the research project using conclusions from the literature review to inform a critical evaluation.

Impact Statement

This research generates insights into the practical application of algorithms in decision-making for mental health outcomes and the development of these systems. The literature review provides a summary of risks from these algorithms and guidelines for improving chances of successful deployment, which has been used in a parliamentary inquiry by the Science and Technology Committee into algorithms in decision-making[1]. The guidance can be used to inform the development of any decision-making system designed to improve mental health outcomes—including those in allied professions, whether from academia, health services, or wider industries.

The methods used to generate these guidelines offer a useful example of a modified realist synthesis of diverse data sources, which can also be applied to disciplines other than psychology. This is a unique design combining both Cochrane and Realist approaches to analyse papers in a data-rich way, which can be useful to future reviewers wanting to explore their data more thoroughly.

The results of the research project will directly inform a future randomised control trial of the main algorithm, which is improved upon in the study. It is hoped the context-specific nature of the algorithm as demonstrated in the research will encourage other studies to consider this as part of their research design, and so generate more reliable results.

The findings and methodology from a case report in Volume Two of this thesis have been presented at two conferences and one user experience design group, and a summary of its research recommendations disseminated as a postcard.

[1] The submission can be found at: http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/science-and-technology-committee/algorithms-in-decisionmaking/written/71708.html

TABLE OF CONTENTS

Part One—Literature Review

Part Two—Empirical Paper

Part Three—Critical Appraisal

Appendices

LIST OF TABLES AND FIGURES

Part One

Part Two

Part Three

Appendices

List of abbreviations

AAU—Allocation as Usual

BIC—Bayesian Information Criterion

CBT—Cognitive Behavioural Therapy

CCG—Clinical Commissioning Group

CI—Confidence Interval

EM—Expectation Maximisation (a form of missing data analysis)

GAD7—Generalised Anxiety Disorder 7 item (questionnaire)

IAPT—Improving Access the Psychological Therapies (primary care mental health service)

IMD—Index of Multiple Deprivation

LP—Latent Profile

mBCH—modified Bolck, Croon and Hagenaars correction (a form of latent profile analysis)

MAP—Maximum a Posteriori (a form of latent profile analysis)

MDS—Minimum Data Set (a questionnaire battery)

ML—Maximum Likelihood correction (a form of latent profile analysis)

OR—Odds Ratio

PCDSS—Predictive Clinical Decision Support System

PHQ9—Patient Health Questionnaire 9 item

PP—Posterior Probabilities; +PP—with the addition of PP (a grouping of models)

RCT—Randomised Control Trial

SP—Sub-Profile

TAU—Treatment as Usual

WSAS—Work and Social Adjustment Scale

The man

Of virtuous soul commands not, nor obeys:

Power, like a desolating pestilence,

Pollutes whate'er it touches, and obedience

Bane of all genius, virtue, freedom, truth,

Makes slaves of men, and, of the human frame,

A mechanised automaton.

Shelley, P. B. (1821; Book III, lines 173-179). *Queen Mab.* London: W. Clark.

Emphasis added.

PART ONE—LITERATURE REVIEW

THE EFFECTIVENESS AND RISKS OF PROSPECTIVE CLINICAL DECISION SUPPORT SYSTEMS IN MENTAL HEALTH:
A SYSTEMATIC REVIEW

1. Abstract

“This data makes a man: A and C and T and G. The alphabet of you, all from four symbols. I’m only two: one and zero.”

“Half as much but twice as elegant, sweetheart.”

(Computer simulation in conversation with a human replicant. Kosover et al., 2017)

Background: Clinicians increasingly use tools to predict who will likely benefit from particular decisions regarding mental health treatment. The performance of these tools in practice—understood as their effectiveness in improving client outcomes, risks from their use, and risks to their uptake by professionals—is not well understood. The purpose of this study is to indicate the feasibility, effectiveness of, and risks associated with prospective clinical decision support systems (PCDSSs) in enhancing client mental health outcomes.

Method and data sources: A systematic review of English language articles using 20 research databases, including PubMed, MEDLINE, and ScienceDirect. Additional studies were identified from experts in the field. Search terms included: decision making, clinical judgement, tool, algorithm, mental health, and practitioner. Pseudo-/controlled quantitative studies were selected to answer the question of clinical effectiveness. Qualitative studies and reported issues in the other selected studies were used to describe risks.

Results and synthesis: Thirty papers met the inclusion criteria. Results were described narratively and combined using a realist synthesis method. A meta-analysis was performed on a subset of three papers on the effect of PCDSSs on assessment (7690 events;

pooled OR 0.93, 95% CI 0.66-1.33) and counselling (879 events; pooled OR 12.62, 95% CI 2.27-70.20) for smoking cessation compared to decision-making as usual.

Conclusions: Results on effectiveness were mixed. A synthesis of the data suggests PCDSSs are more likely to have a positive impact when they function by working collaboratively with the clinician, and appropriately integrating research, expert and contextual evidence to form the best 'ecological fit' between tool, practitioner, organisation, and client. In practice this is rarely the case, meaning PCDSSs are less likely to be trusted, are liable to rejection by clinicians, and have difficulty accounting for organisational and client contexts. Feasibility and effectiveness can be 'designed in' with appropriate stakeholder involvement.

2. Introduction

> "I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted."
>
> (Turing, 1950, p. 442)

And Turing was right: we believe machines think. In fact we are now simply planning for the machine-driven healthcare revolution (Chan, 2017). Advances in computer power, access to data, and decisional modelling saw computers mimic human thought well before the end of the 20th Century, and in some respects surpass it by 1997 when Deep Blue famously defeated the world chess grandmaster, Garry Kasparov (IBM, 2015). While progress has been fast, using simple algorithms to assist clinical decisions has long been commonplace: the Hospital Anxiety and Depression Scale is 30 years old (Zigmond & Snaith, 1983) and the first formal personality tests appeared 100 years ago (Woodworth, 1917); today there are more than 250,000 healthcare algorithms in the literature (Iyengar, 2009). The spread of information technologies has allowed these processes to be automated in routine care, and the rise of Big Data promises a new Deep Blue for mental health judgements. However, as chess grandmasters have not yet been replaced by computers, neither have clinicians.

This is rather curious, as clinicians (and people generally) have long been recognised as poorer decision makers compared to their algorithmic counterparts (for example Meehl, 1954; Clark, 1992), and this does not appear to have improved with time (Ægisdóttir et al., 2006). Humans are particularly bad at judgements under conditions of uncertainty (Tversky & Kahneman, 1973), such as predicting the future, and are consistently outperformed by statistical models—even when more data is available to the human (Meehl, 1954; 1986).

Greater information may even lead to worse judgements (Bastardi & Shafir, 1998), making the impending data revolution a cause for concern. Yet on the other hand there are an increasing number of cases where algorithms have disastrous errors in judgement, from the inadvertently racist and sexist to wiping billions from the stock market based on a tweet (e.g. Karppi & Crawford, 2016; Robb, 2017; Simonite, 2017). Are then clinical decision support systems as effective as they are assumed to be, given that there are significant risks involved in their use and they may impair human decisional capacity overall? Is this why they not been adopted more widely?

The latter question taps more broadly into the issue of implementing evidence-based care. There exists a 'translational gap' between research recommendations and practice (Hemmelgarn, Glisson, & James, 2006; Lenfant, 2003), and of those guidelines that do make it into services relatively few lead to sustained improvement (Scheirer, 2005; Shortell, Bennett, & Byck, 1998). Development of guidelines alone is not enough to create change (Lugtenberg, Burgers, & Westert, 2009), and not addressing barriers to adoption risks wasting resources, duplicating work, and reducing confidence in similar endeavours. Therefore it becomes imperative to understand both the risks *from* decision tools and *to* their implementation in order to understand findings on effectiveness. And, as with the best narratives on artificial intelligence, threats often come from both human- and computer-based systems.

2.1. Predictive Clinical Decision Support Systems

PCDSSs are defined here as algorithms that anticipate how a given individual will respond in a particular clinical situation, and make a recommendation for action based on this. Individual data of some form, such as diagnoses, number of clinical symptoms, or demographics, are processed according to a set of rules that result in a prediction of which intervention (or not) is most likely to produce beneficial change in that individual. In the cases studied here a clinician is required to make the final decision for care, so the prediction is always a recommendation for action. A simple PCDSS is given in Appendix B based on a standard depression questionnaire and gives evidence-based recommendations depending on the score; however PCDSSs can be more complex than this. Some generate specific probabilities for outcomes, others involve continuous learning and artificial intelligence, for instance.

Since algorithms themselves are sets of rules in a calculation, PCDSSs are problem-solving models. This makes algorithms useful for understanding how clinicians make decisions (e.g. Kahneman & Tversky, 1979), and how to make better clinical decisions (Grove et al., 2000). Today's access to large amounts of data, partly driven by the rise of electronic health records, makes accurately predicting outcomes more viable than at any point in history. PCDSSs have appeared that identify clients at risk of treatment failure (Hannan et al., 2005), children who might develop antisocial personality disorder (Lahey, Loeber, Burke, & Applegate, 2005), and parents who may maltreat their child (Bugental & Happaney, 2004), to name but a few. Given the particular human difficulty with judgements under uncertainty, designing systems to support clinical predictions could significantly improve treatment for existing problems, and prevent others from developing at all.

Most of the literature on PCDSSs for healthcare focuses on medication or physical health. For instance, a systematic review by Adli, Bauer, and Rush (2006) of collaborative-care algorithms for depression found treatment outcomes using antidepressants were improved through changes to practice and treatment procedures. Further improvement was limited by the available evidence on pharmacogenetic predictors of response. In Garg et al. (2005), which included mental health diagnoses in its outcomes, clinician behaviour was positively affected by PCDSSs when they were automatically prompted to action, and were involved in the design of the system. There is little evidence on algorithms as an aid to decision-making for mental health outcomes in general as might be applicable to clinical psychology or similar psychologically-informed practice.

There is scarce collected research on the risks from algorithms in clinical practice. In physical medicine, several studies have highlighted problems related to knowing how algorithms process data. In Rajkomar et al. (2018), machine learning (algorithms that perform a given task without being given explicit instructions for how to do it, thereby generating their own models instead) was used predict medical events for hospitalised patients. The system outperformed traditional clinical models in all measures, however clinicians were unable to fully understand the new predictive models. Burt and Volchenboum (2018) point out that if clinicians do not understand how a tool works, they cannot correct errors it might make. If these systems are then used, new and unknown dangers are likely to emerge. The author is unaware of similar research on the risks of algorithms in psychological practice.

2.2. Research questions

This research seeks to explore the effectiveness of predictive clinical decision support systems (PCDSSs) to improve mental health outcomes, including their ability to affect these changes in practice. A PCDSS is understood as any rule-based system involving calculations performed by a machine (or could feasibly be done so) in order to prospectively determine who can benefit from a particular intervention. This can include recommendations for therapy based on analysis of wellbeing scores, suggestions for further assessment after a new diagnosis, and other such proactive measures based on a prediction of client responsiveness.

This review will attempt to answer:

1. What is the evidence for the use of prospective tools in clinical decision-making to improve mental health outcomes in clients?
2. Is the use of prospective clinical decision support systems (PCDSS) feasible in mental health practice?
3. What are the risks associated with PCDSSs in mental health settings?

To the author's knowledge, a systematic literature review of PCDSS effectiveness and risks in mental health settings has not so far been attempted, and would be beneficial in light of current trends in the use, misuse, and under-use of data.

3. Method

Due to the specialist nature of the subject, the methodology was designed to cast as wide a net a possible for potentially relevant research to increase the chances of finding a usefully large number of studies. This allowed the resulting analysis to comment broadly on the usefulness of PCDSSs in mental health rather than on a more specific clinical area.

3.1. Protocol

Methods for the search strategy and inclusion criteria were specified before searching was commenced in order to reduce potential bias. The full protocol can be found in Appendix C.

3.2. Eligibility criteria

3.2.1. Studies

To be eligible for inclusion, the research must have demonstrated it tested the effectiveness of a PCDSS. Quantitative experimental or quasi-experimental designs only were included to answer Question One. No design exclusions were applied for other research questions. Only studies that could be accessed and read by the author were included, which was limited to

English language publications available through the UCL library. Unpublished research and theses were also included. No books were included in the search scope as these were less likely to contain new research and would be significantly harder to search. Publication years were not specified, i.e. searches were from the earliest date in the database to 2017.

3.2.2. Participants

Those using the PCDSS must have been healthcare staff, and the PCDSS must have been applied to clients in a healthcare setting.

3.2.3. Interventions

Studies must have used a tool that was used to prospectively assess clients, that specifically informed clinical decision-making, and did this (or could feasibly do this) automatically. This did not have to be in a mental health setting, but could reasonably be expected to be used in such a context.

3.2.4. Outcome measures

At least one outcome must have been relevant to a mental health setting. This included any measure that directly involved psychological processes, such as anxiety symptoms, adherence to treatment, or improvement in quality of life (when psychological factors are taken into account). Also included were any outcomes detailed in the Public Health England (2014) Priorities report, such as smoking reduction or improvements in alcohol dependency.

3.3. Information sources

Studies were identified by searching databases and consulting experts in the field of mental health decision support algorithms. Twenty seven databases were identified as potentially useful to search. The databases searched to completion were: ASSIA Applied Social Sciences, CINAHL Plus, Health and Psychosocial Instruments, International Bibliography of the Social Sciences, IngentaConnect, Journals@Ovid, MEDLINE, Psychoanalytic Electronic Publishing, Published International Literature On Traumatic Stress, ProQuest Central, Psyc-ARTICLES, -EXTRA, -INFO, and -TESTS, Pubget, PubMed, Science Citation Index Expanded (Web of Science), ScienceDirect (Elsevier), SCOPUS, and University of London Research Library Services.

3.4. Search

The following search string was used to examine all databases. The search string was trialled before use with studies identified previously to check sensitivity. Earlier iterations are documented in the Protocol. Filters for full text availability and removal of results attached to non-human index terms were used where permitted. Where it was not possible to specify certain search areas (such as the Title, Abstract, etc.) for a given database, predefined substitutions were used.

Search: Title and Abstract ((“decision making” OR “decision-making” OR “clinical judgement” OR “clinical decision” OR predict OR “care suggestions” OR “care process” OR “care processes”) **AND** (tool OR “decision support” OR “decision rules” OR algorithm OR aid OR “care suggestions” OR “treatment advice”) **AND** (“mental health” OR “mental illness” OR wellbeing OR depression OR anxiety OR admission OR discharge OR referral OR “quality of life” OR “treatment response” OR “response to treatment”) **AND** (psychologist OR psychologists OR professional OR professionals OR clinician OR clinicians OR practitioner OR practitioners OR provider OR providers OR physician OR physicians) **NOT** “shared decision”))
Search: Title (**NOT** (“systematic literature” OR “systematic review”))

3.5. Study selection

Eligibility assessment was standardised and performed by the author according to the Protocol. Reliability was tested using a senior researcher independent to the study, who used the same procedures on a subset of databases. Disagreements were resolved by consensus and the Protocol updated accordingly.

Records were retrieved using the search string, and titles screened for potential relevance. Abstracts for these studies were then reviewed, creating a short-list of studies for full-text assessment. Databases were examined sequentially, and duplicates were removed at this stage before attempting to access complete texts. The eligibility criteria were used to determine whether a study qualified for selection.

Studies were selected based on their coverage of a PCDSS. Where a paper indicated other publications reported on different aspects of a single PCDSS as part of the same overall study, such as separate papers on qualitative and quantitative findings, these were identified manually through citations searches.

3.6. Data collection process

Data from studies were recorded in a pre-defined data extraction form (see Protocol, Section 6.a). A random subset of studies was used in the validation procedure, where an independent researcher used the form to record data items. This was checked for consistency with the first author.

3.7. Risk of bias

For individual studies, quality was assessed using the Mixed Methods Appraisal Tool (MMAT; Pluye, Robert, Cargo, & Bartlett, 2011; see Appendix D), a brief checklist of important criteria for specific designs. For mixed methodologies, criteria are combined from individual designs such as qualitative and descriptive. Studies receive one point for each criteria reported, for example the criterion ‘60% or higher response rate’, for a maximum of four points. The MMAT was chosen for its ability to indicate methodological rigour for a variety of study designs in a single tool.

Risk of bias across studies was assessed with an adapted form of the Cochrane Collaboration tool (Higgins & Green, 2011) using applicable MMAT scores from individual studies and the author’s judgement to gauge overall bias.

3.8. Methods of analysis

Studies were divided into two broad groups for analysis representing PCDSS effectiveness in practice and risks. Controlled and pseudo-controlled quantitative studies were used for research questions One and Two on the effectiveness and feasibility of PCDSSs. All other studies, including qualitative reports within the controlled studies, were collated to investigate research Question Three on the risks PCDSSs posed. A specific method of analysis was not pre-defined in the protocol as the type of data available would not be known until the literature search was complete. Where more than one study reported on a single PCDSS these were treated as one study for the analysis, while MMAT scores were given separately.

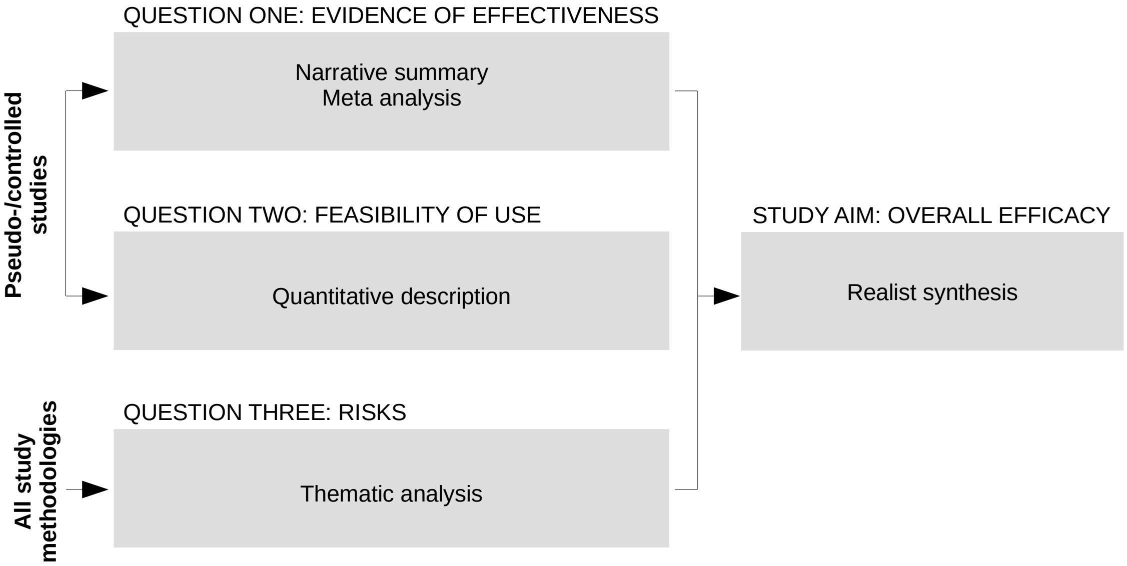


Figure 1: Summary of analytic methods for each research question

On reviewing the types of evidence found, five methods of analysis were chosen to provide meaningful summaries within the context of large heterogeneity, as summarised above in Figure 1. For the narrative summary, quantitative findings were summarised for each eligible study. A meta-analysis on a sub-set of these papers with similar designs was conducted to estimate a pooled effect size. This used a random effects model to account for heterogeneity between studies. For an estimate of feasibility, numerical data on assessment rates and adherence to PCDSS recommendations were recorded in a table against established feasibility indicators (Benbenishty & Treistman, 1998; Cooley et al., 2015). For Question Three (risks) all papers were read in detail and relevant qualitative information extracted. Individual items were then iteratively organised into common thematic areas, and these findings narratively summarised. Textual descriptions were chosen to give a preliminary overview of the findings (Snilstveit, Oliver, & Vojtkova, 2012). A modified realist synthesis was performed on the entire data set to understand the efficacy of PCDSSs more completely (Pawson, Greenhalgh, Harvey, & Walshe, 2004). This synthesis was chosen for its ability to reflect the complexity of healthcare interventions, which include variable designs, implementations, management, and service regulations (Pawson et al., 2004). This method is particularly sympathetic to the analysis of diverse methodological and multidisciplinary approaches found in this review, and is provided in more detail next.

3.8.1. Realist synthesis method

The aim of a realist synthesis is to develop a model of how context, agents, and circumstance mediate the outcome of an intervention. Briefly, this is achieved by pre-defining the 'theories' thought to underpin the effectiveness of an intervention ('why is it thought to work?') and to interrogate these using the selected studies ('how does it work in practice?'). In the process, conditions under which the intervention is effective or not are identified, particularly through results that may contradict other findings ('why does it work here and not there?'). The resulting synthesis should answer the question 'when does the intervention work?', providing a richer understanding of the data compared to 'does it work?'.

This review broadly followed the method described by Rycroft-Malone et al. (2012). First, theories (implicit assumptions of how an intervention brings change) underlying the use of PCDSSs were defined. Data from the preceding qualitative and quantitative analyses were then organised according to the theory they provided evidence on. These evidence clusters were coded to create themes related to that theory, with particular attention paid to contradictions in the data. Contradictions and challenges were used to suggest conditions under which a PCDSS was un/successful, thereby building chains of inference from individual themes. Chains were then related back to the original theory and theory areas to create hypotheses on the conditions under which PCDSS were in/effective. Analysis from themes to hypotheses was iterative, and these steps were repeated until a coherent picture of the data was created. In order to reduce the risk of selection bias all identified studies were included in the analysis, and can be directly traced back from hypotheses to theories (this can be done using Appendix E and Appendix F). Lastly, hypotheses were expanded on as narratives.

Realist syntheses usually include a review of findings with commissioners and other stakeholders to complement hypotheses with additional knowledge not available in the searched literature (Pawson et al., 2004; Rycroft-Malone et al., 2012). This was not possible for the present study, so an alternative step was designed specifically for this research to maintain higher, if not ideal, standards of validity. This ‘expert consultation’ involved checking hypotheses against known literature on decision-making to fill knowledge gaps not otherwise addressed by the studies identified. This was to produce a fuller explanation for the data as well as cross-check hypotheses for potential short-comings.

4. Results

Of the 27 databases identified, 20 were searched. Six could not be used as their search engines were not complex enough to handle the pre-defined search string. One database (Wiley Online Library) was excluded from further searching after producing no relevant results in the first 100 entries. Twenty four studies were found from eight databases and two identified before searching began, representing 26 separate studies across 30 papers. The search process is illustrated below in Figure 2.

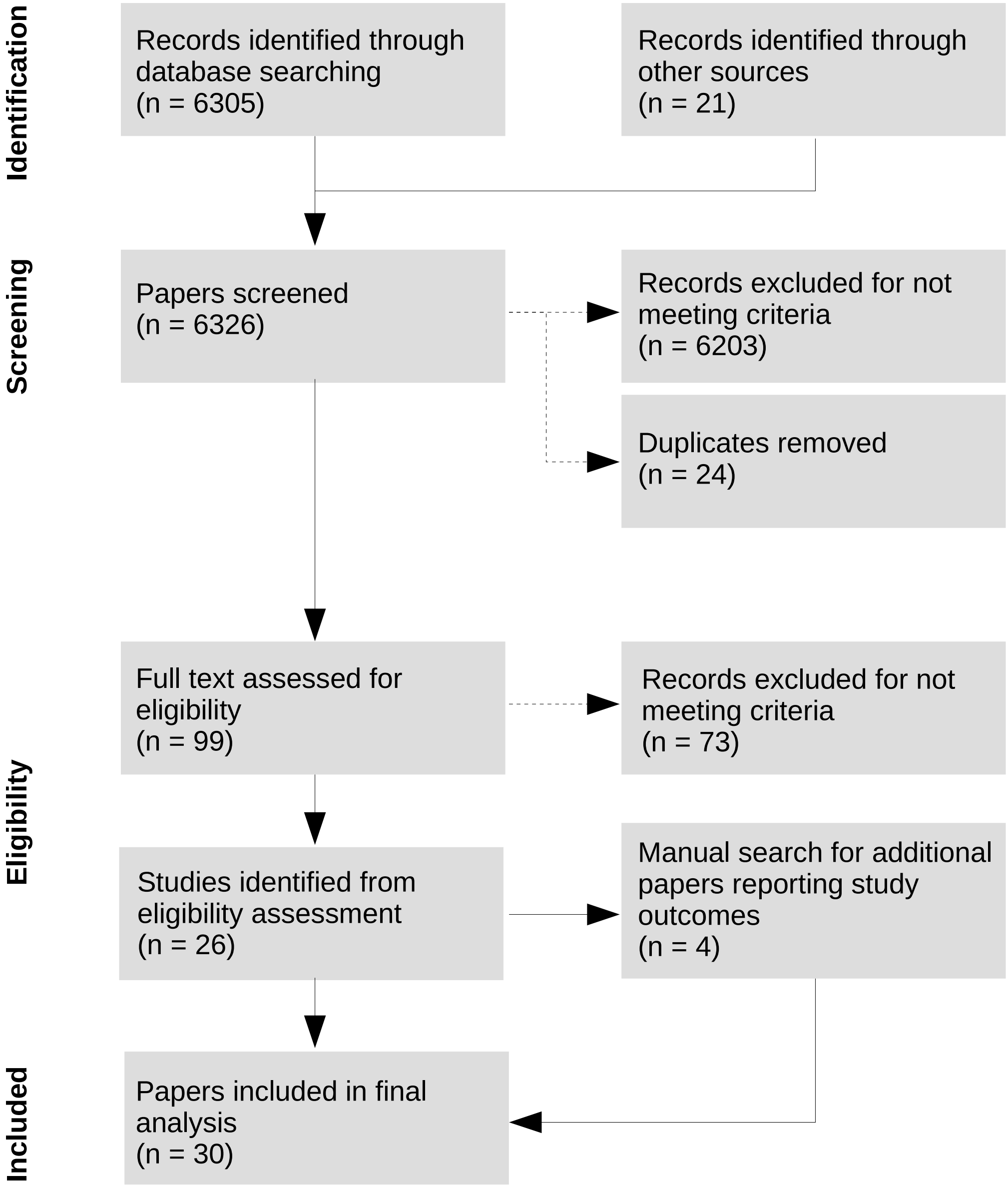

Identification
Records identified through database searching (n = 6305)
Records identified through other sources (n = 21)
Screening
Papers screened (n = 6326)
Records excluded for not meeting criteria (n = 6203)
Duplicates removed (n = 24)
Eligibility
Full text assessed for eligibility (n = 99)
Records excluded for not meeting criteria (n = 73)
Studies identified from eligibility assessment (n = 26)
Manual search for additional papers reporting study outcomes (n = 4)
Included
Papers included in final analysis (n = 30)

Figure 2: Flow of information through stages of the systematic review, based on (Moher, Liberati, Tetzlaff, & Altman, 2009)

The studies totalled 11 qualitative designs, eight randomised control trials, five pseudo-controlled studies (non-randomised designs with a comparison group), five mixed designs, and one descriptive, involving at least 25,177 clients and 805 healthcare professionals. Eight studies were set in primary care clinics, six in mental health facilities, two in general hospitals or paediatric clinics, and eight in various specialist settings. The most frequent intervention target was common mental health issues, such as depression or anxiety. Studies and their outcomes are summarised in Table 1 as follows:

Table 1:
Summary of research papers included in the final analysis, with findings and quality scores

Study reference(s)	Design	Setting	Intervention target area	Impact of PCDSS	MMAT*	Sample size
Carroll, Biondich, Anand, Dugan, & Downs, 2013a	RCT	Paediatric clinic	Maternal depression	Higher rate of detection but lower referrals with PCDSS. Referrals occurred earlier	1	3520 children; 48 physicians
Carroll et al., 2013b		Primary care clinic	Childhood ADHD	Higher use of structured diagnostic assessment and more symptoms recorded. No change to treatment decisions	3	84 service users; unknown number of clinicians
Huijbregts et al., 2013		Primary care clinic	Depression	Larger and faster response to treatment. Higher dropout compared to control	1	152 service users; 82 GPs
Rindal et al., 2013		Dental clinic	Tobacco smoking	Increase in assessment and referrals for smoking cessation	1	579 service users; unknown number of clinicians
Rollman et al., 2002		Primary care clinic	Depression	No significant change to detection or treatment	3	8302 service users; 17 GPs
Thomas et al., 2004		Primary care clinic	Anxiety and Depression	Greater initial response to treatment, not maintained at 6 months. No change to quality of life or satisfaction with treatment	2	762 service users; unknown number of clinicians

Tolin, Diefenbach, & Gilliam, 2011		Mental health service	Obsessive Compulsive Disorder	No change in response. Lower treatment costs in PCDSS group	0	30 service users; unknown number of clinicians
Chorpita et al., 2007; Weisz et al., 2012	RCT; qualitative	Mental health service	Child anxiety, depression and conduct problems	By end of treatment, fewer problems reported by both children and parents and less diagnoses. Faster response to treatment. Less time in treatment compared to control	4; 4	174 children; 84 therapists
Bowles, Hanlon, Holland, Potashnik, & Topaz, 2014	Pseudo-control	General hospital acute ward	Older adult readmission within 60 days	Fewer readmissions for high risk patients	3	533 service users; unknown number of clinicians
Clarke, Brown, & Griffith, 2010		Psychiatric inpatient facility	Inpatient violence	Reduction in use of seclusion to control violence. Potentially lower rates of violence overall	3	277 service users; 54 healthcare professionals
Sharifi et al., 2014		Paediatric clinic	Tobacco smoking in parents	No change to screening rates. Increased counselling and quit referrals for positive screens, maintained at 9 months	3	3919 service user contacts; 48 clinicians

Stallvik, Gastfriend, & Nordahl, 2015		Substance use treatment centre	Treatment allocation for substance use	Matching patients to guideline recommendations reduced substance use severity scores, improved retention, and increased the proportion ready to step down to a lower level of treatment	3	261 service user cases; unknown number of clinicians
Foster et al., 2014; Sanders, Foster, & Ong, 2011	Pseudo-control; qualitative	Primary care clinic	Lower back pain	Improved fear and avoidance beliefs, and less time away from work. Reduced average cost of treatment and increased quality of life scores	2; 4	922 service users; 41 clinicians
Jenssen et al., 2016	Mixed: pseudo-control and qualitative	Primary care clinic	Tobacco smoking	Reduction in number of people screened compared to baseline. Increased number of people offered treatment and range of therapies prescribed	3	200 service users; 30 clinicians
Olfson, Tobin, Cassells, & Weissman, 2003		Primary care clinic	Substance use and depression	No change to whether condition was recognised and discussed by clinician, a mental health referral made, or treatment prescribed	3	467 service users; 11 clinicians
Benbenishty & Treistman, 1998	Mixed: descriptive and qualitative	Military clinic	Discharge from the military on mental health grounds	PCDSS agreed with clinician decisions more than clinicians agreed with each other. System generally not acceptable to clinicians	4	52 service user cases, 6 mental health officers

Kennedy, Finkelstein, Hutchins, & Mahoney, 2004		Perinatal clinic	Prenatal maternal substance use	Increase in detection of substance use and delivery of brief intervention or referral to specialist	3	4660 service users, 175 healthcare professionals
Cooley et al., 2013, 2015; Lobach et al., 2016	Mixed: descriptive and qualitative; descriptive; qualitative	Outpatient cancer facility	Pain, fatigue, depression, and anxiety in thoracic cancer	Assessment and adherence to recommendations was below level of feasibility. System generally acceptable	3; 4; 3	145 service users; 32 professionals
Barnett, dosReis, & Riddle, 2002	Qualitative	Residential youth psychiatric facility	Inpatient aggression	System generally acceptable to professionals	3	42 healthcare professionals
Buckingham et al., 2015		Mental health service	Risk of self harm, suicide, vulnerability, self-neglect, and harm to others	System generally acceptable to clinicians and service users. Improved clinicians' perception of their ability to explain judgements to clients	3	115 service users; 93 clinicians
Chase, 2014		Hospital inpatient ward	Clients with impaired decision-making capacity	Improved staff understanding of capacity issues, decreased anxiety about identification and management, and improved teamwork with safety officers	3	Unstated number of healthcare experts

Colombet et al., 2003	Primary care clinic	Depression and suicide, as well as cardiovascular problems and preventable cancers	Poor adherence by clinicians to recommendations	3	11 GPs; unknown number of service users
Hunter et al., 2016	Local commissioning body	Setting priorities in public health spending	No significant changes; overall rated positively by professionals	3	29 professionals; 19 expert users
Nagpaul, 2001	Older adult services	Elder abuse	Evidence of improvement in detecting abuse and taking appropriate action	2	7 case studies of service users and clinicians
van Vliet, Harding, Bausewein, Payne, & Higginson, 2015	Palliative care service	Family anxiety, depression, breathlessness and information needs	System generally acceptable to practitioners and service users	4	4 service users/carers; 34 professionals
Wilkinson & Himstedt, 2008	Mental health services	Malnutrition	System generally acceptable to practitioners	2	Unstated number of healthcare professionals

*Score is out of four, where four indicates a high design standard

4.1. Risk of bias in included studies

One study scored zero on the MMAT, but this was judged to not accurately reflect its design standard, and was not excluded from analyses. Overall risk of bias was assessed based on applicable MMAT criteria and aggregated across included studies. This is summarised in Figure 3, below. Overall risk of bias was judged to be low to moderate.

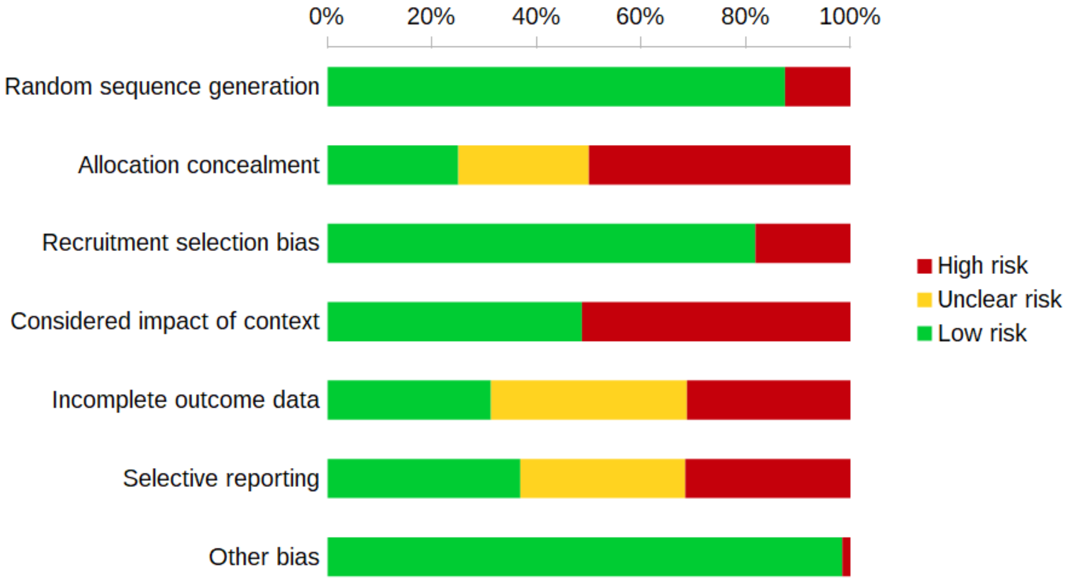


Figure 3: Risk of bias graph—author's judgement on risk of bias for selected criteria summarised across included studies. Adapted from Higgins and Green (2011)

RCTs were generally good at describing their process of randomisation. However, due to the practical difficulty of blinding clinicians to PCDSS use, concealment was rare and sporadically reported for patients. One study experienced contamination between intervention arms, but not with the control arm, suggesting quantitative findings were likely biased, although the qualitative discussion was still useful for the analysis.

Few studies considered the potential impact of context or the researcher on findings, such as the influence of client demographics, service attitudes, or motivations for research.

Where this was recorded it was generally not reported in any depth. For the present research, contextual factors were examined and cross-referenced between studies and the literature as part of the analysis on conditions of effectiveness—as even when absent from research discussions it was still possible to draw inferences. This would likely reduce the impact of this bias on the overall findings.

Several studies suffered from selective reporting, here more associated with a high drop-out rate or misleading statistics. For instance one study claimed a higher rate of referral for its PCDSS based on absolute values, yet when relative proportions were compared it was significantly lower than baseline. This has been corrected for in the current analysis where observed. It has usually not been possible to determine the circumstances under which participants dropped out of studies. This suggests conclusions on service users may be biased, potentially towards those more readily agreeable to PCDSSs.

4.2. Research Question One: What is the evidence for the use of PCDSS in clinical decision making in mental health?

PCDSSs were most commonly used to improve assessment rates for problems and allocate treatment for clients screening positive. Evidence of efficacy was mixed, with some tools significantly improving identification and referrals, while others produced no measurable difference in clinician behaviour or outcomes. This is analysed further in the synthesis (see 4.6).

4.2.1. Health psychology

Foster et al. (2014) found a PCDSS led to a significant increase in appropriate referrals for back pain in medium-to-high risk clients (from 40% to 72%), with associated improvement in disability (average RMDQ change of 0.7), quality of life (QALY 0.003-0.008), pain, fear and avoidance beliefs, time off work (30% reduction in sickness certification; 50% less time off work), and depression. Savings were approximately £34 per client, or £400 per employed client per 6 months. No significant change was found for low-risk clients.

Bowles et al. (2014) found using a PCDSS to examine risk factors including mental and emotional needs reduced hospital readmissions for older adults at 30 and 60 days after discharge from a medical unit, representing a 26% relative reduction overall.

4.2.2. Substance use

A PCDSS did not affect rate of recognition or treatment of substance use by primary care practitioners, even with feedback on positive screens (Olfson et al., 2003). However, in a specialist treatment facility, Stallvik et al. (2015) found matching clients to guideline recommended care resulted in significantly less alcohol and cannabis use, and greater retention (62%) within the treatment programme compared to 'under-matched' controls (45%). 'Over-matched' clients had similar outcomes to matched users. Severity across multiple domains was also more likely to fall, and readiness to step-down level of care increased (61% compared to 46% under-matched and 17% over-matched).

Several studies investigated the impact of PCDSSs on assessment and treatment for smoking cessation in non-smoking-related healthcare facilities. Assessment results were mixed: Sharifi et al. (2014) found no change in assessment by clinicians, Rindal et al. (2013) identified a moderate improvement (70% for control and 87% for the intervention condition), while Jenssen et al. (2016) found a small decrease compared to baseline (from 82% to 76%). However, where assessment screening was positive, PCDSS use was linked to an increase in the rate of cessation counselling and referral. These results are examined further in the meta-analysis that follows (section 4.3). Only Jenssen reported on service user uptake following referral, which was zero from 165 referrals.

4.2.3. Common mental health conditions

Using a PCDSS had no effect on general practitioners' recognition of depression in Olfson et al. (2003), and neither Olfson nor Rollman et al. (2002), who used automatic screening, detected a difference in treatment offered. This is in contrast to Carroll et al. (2013a), where automated screening significantly increased the rate of identification of maternal depression, compared to a control group that was only reminded to screen. However, referral rates dropped overall in the Carroll study from 100% of identified cases to 28%. This could be explained if only severe cases were identified by clinicians in the control group, compared to a wider spectrum in the intervention that did not necessarily all require referral. The results from Olfson and Rollman may be understood in terms of a lack of effect of the PCDSS on clinician behaviour. Additionally, despite a small increase in the average number of clinician visits in the intervention group, Rollman found depression scores were unaffected by the PCDSS.

Huijbregts et al. (2013) on the other hand found intervention and control conditions received approximately the same amount of care, although the type of contact differed slightly. For example, groups using the PCDSS were more likely to see a mental health practice nurse or social worker and less likely to be admitted to psychiatric facilities. The intervention group had a significantly increased response to treatment at nine months compared to TAU (OR 5.6, CI: 1.40-22.58), had higher rates of remission at 12 months (20.7% intervention versus 6.3% control), and was associated with a faster response time. This suggests the effect observed with the PCDSS may have come from the provision of more effective care, rather than more care per se. Giving information on diagnostic status alone to clinicians is unlikely to produce change, as neither Rollman (2002), Olfson (2003), nor Thomas et al. (2004) found a significant impact on clinical outcomes when PCDSSs only passed on diagnostic data.

Carroll et al. (2013b) and Weisz et al. (2012) found the impact of PCDSSs was enhanced when specific 'modules' were used to target particular problem areas, compared to general protocol-driven procedures. In the Carroll (2013b) study, ADHD-specific modules increased the use of structured diagnostic assessments and the number of ADHD symptoms recorded at time of diagnosis. No significant changes were noted in ADHD care management, although the study was underpowered to detect the expected effect. Trend data indicate the ADHD module may have increased the number of medication adjustments, symptoms reassessments, and mental health referrals. In Weisz et al. (2012) therapy planned using the PCDSS performed significantly better than usual care or protocol-based therapy for children with anxiety, depression, and conduct problems. PCDSS recommendations outperformed usual care and non-modular protocol treatment on both parent and child reported measures, with effect sizes between 0.50 and 0.72. Children treated with the aid of the PCDSS also had significantly faster improvement, fewer diagnoses at the end of treatment, and spent an average of 75 days less in treatment versus usual care. In counterpoint, Tolin et al. (2011) found a protocol-driven stepped care intervention based on a PCDSS decision-tree led to no significant differences in OCD symptoms compared to standard treatment, yet was significantly cheaper.

4.2.4. Violence

Completing a short checklist-based PCDSS assessing risk factors associated with violence reduced the use of seclusion to manage inpatient aggression, with findings sustained at one- and five-year follow-ups (Bowles et al., 2014).

4.3. Meta-analysis on PCDSS effectiveness in tobacco cessation

A meta-analysis was conducted with Review Manager (RevMan, 2014) on three studies (Jenssen et al., 2016; Rollman et al., 2002; Sharifi et al., 2014) to provide an estimate of overall treatment effect for PCDSSs on assessment and counselling in tobacco cessation. These were the only studies with sufficiently similar designs and outcome measures to allow for a meta-analysis. As outcomes were dichotomous, a Mantel-Haenszel estimate was used. Heterogeneity was high, therefore a random-effects model was assumed.

4.3.1. Assessment for tobacco use

Only one study found the use of a PCDSS improved clinicians' rate of assessment for smoking in their patients. The pooled analysis suggested PCDSSs overall are not effective at increasing screening for tobacco use, as shown next in Table 2 and Figure 4.

Table 2
Clients assessed for tobacco use

Study	**PCDSS**		**Control**		**Weight**	**Odds Ratio [95%CI]**
	Events	Total	Events	Total		
Rindal et al. (2013)	200	263	195	285	29.1%	1.47 [1.00, 2.14]
Sharifi et al. (2014)	684	2024	719	1895	41.2%	0.83 [0.73, 0.95]
Jenssen et al. (2016)	2286	3023	163	200	29.7%	0.70 [0.46, 1.02]
Total [95% CI]		5310		2380	100.0%	0.93 [0.66, 1.33]

Heterogeneity: $\text{Chi}^2 = 8.98$, df = 2, p=0.01, $I^2 = 78\%$
Test for overall effect: Z = 2.244, p=0.01

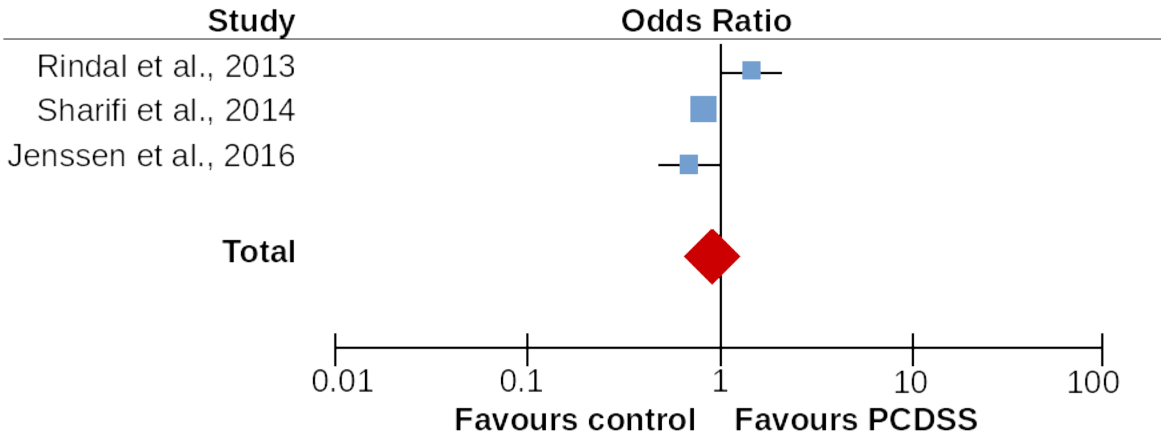


Figure 4: Forest plot of odds ratios for clients assessed for tobacco use. This figure suggests PCDSSs may be not considered effective at improving assessment among clinicians

4.3.2. Provision of counselling

Where assessment for smoking occurred, all three studies found the PCDSS was superior to control in prompting clinicians to offer advice on cessation (see Table 3 and Figure 5). The pooled effect size of 12.62 is likely to be an overestimate of the difference due to an unusually high odds ratio from the Jenssen study, which may be an unreliable indicator of relationship strength as suggested by the large confidence interval. However, the direction of the effect was the same for all studies.

Table 3
Clients counselled to stop smoking

Study	**PCDSS**		**Control**		**Weight**	**Odds Ratio [95%CI]**
	Events	Total	Events	Total		
Rindal et al. (2013)	123	263	73	285	36.2%	2.55 [1.78, 3.66]
Sharifi et al. (2014)	67	112	13	117	34.7%	11.91 [5.98, 23.73]
Jenssen et al. (2016)	66	69	6	33	29.1%	99.00 [23.07, 424.77]
Total [95% CI]		444		435	100.0%	12.62 [2.27, 70.20]

Heterogeneity: Chi^2 = 34.47, df = 2, p<0.001, I^2 = 94%
Test for overall effect: Z = 9.98, p<0.001

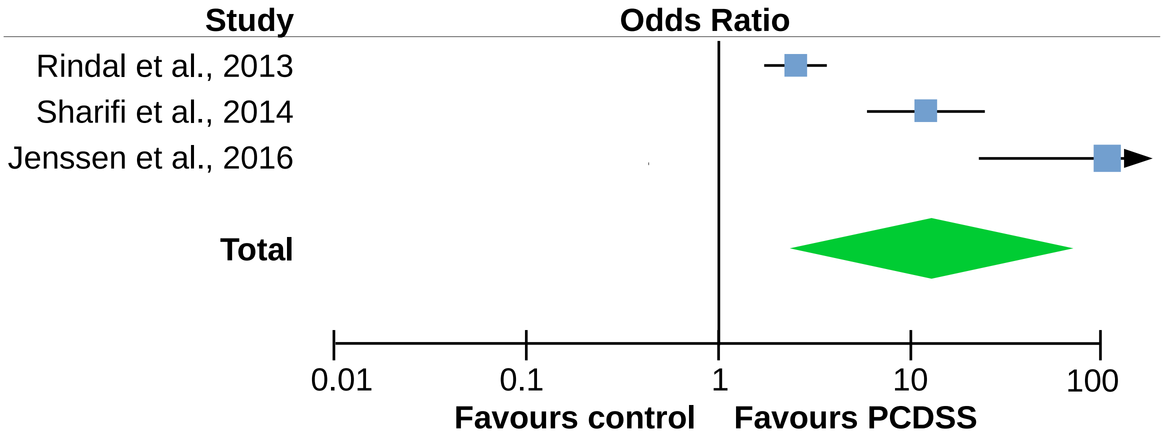


Figure 5: Forest plot of odds ratios for clients counselled to stop smoking. This figure suggests PCDSSs may be considered effective at improving counselling rates

4.4. Research Question Two: Is the use of PCDSSs feasible?

Nine studies reported the proportions of clinicians using the PCDSSs for assessment and adhering to recommendations in practice (see Table 4). Judgement of feasibility is based on Cooley (2013) and Benbenishty and Treistman (1998). Benbenishty argues 75% agreement with recommendations is an appropriate benchmark as it is marginally better than the 70% inter-practitioner reliability indicated by previous studies. 'Feasibility' was therefore defined as: at least 75% of clinicians using the PCDSS to assess and (additionally) at least 75% of clinicians fully adhering to those recommendations. In some papers assessment was performed by the researchers, in which cases judgement on whether practitioners would assess unaided is unclear. In these cases feasibility is judged to be 'possible', as long as the adherence to recommendation criteria is met.

Table 4
Percentage of cases where PCDSS recommendations for assessment or treatment were adhered to

Study	**Percentage of eligible assessments made with PCDSS**	**Percentage recommendation adherence**	**Feasible?**
(Cooley et al., 2015)	84	57	No
(Kennedy et al., 2004)	95	77	Yes
(Olfson et al., 2003)	100*	48.3 (partial adherence)	No
(Sharifi et al., 2014)	36.1	67	No
(Rindal et al., 2013)	87	74	No
(Huijbregts et al., 2013)	100*	90.9 (partial adherence)	Possibly
(Benbenishty & Treistman, 1998)	100* (study indicates this is closer to 15% without researcher input)	86.5	Unlikely
(Carroll et al., 2013b)	81	No significant differences between groups	Possibly
(Clarke et al., 2010)	76	64	No

*assessment completed by researchers

Only one study clearly indicated PCDSS use was feasible in practice based on these criteria. Overall it seems unlikely PCDSSs would be effectively deployed to sustainably improve outcomes related to mental health under the conditions described in these studies. Conditions of effectiveness are explored more in section 4.6.

4.5. Research Question Three: What are the risks associated with PCDSSs?

Data used to answer this question are split into risks from PCDSSs themselves and risks to PCDSS uptake. They are organised thematically within each section. Risks from algorithms were associated more strongly with inappropriate recommendations due to limited evidence, such as on individual patient preferences, organisation priorities, or contraindications for therapy. Risks to uptake were linked more to clinician perceptions of low utility or poor organisational fit.

4.5.1. Risks from PCDSSs

1) **Patient-treatment compatibility** (Jenssen et al., 2016; Nagpaul, 2001; Sanders et al., 2011)

   Algorithms are only as good as the data put into them, and research is limited on what works for whom. PCDSSs are therefore less able to match treatment to individual client preferences than they are to diagnoses. For instance, Jenssen's (2016) PCDSS recommended 165 people for a guideline-based tobacco-cessation program, yet none attended. Algorithms also have difficulty allowing for circumstances that break their rules (often the subject of science fiction endeavours). In one example, a tool may always recommend moving victims away from an abuser. However, in cases of abuse by a caregiver, this may be counter to the wishes of the client, and lead to additional distress. Further, the perpetrator themselves may have needs that are not addressed by the PCDSS.

2) **Competing interests** (Barnett et al., 2002; Colombet et al., 2003; Sanders et al., 2011)

   As the purpose of a PCDSS is to alter clinical practice, it will at the very least challenge existing procedures. Since many services operate under constrained resources, a tool is likely to compete for those resources with other interests:

   "Yes, almost everybody with back pain will benefit from that [recommendation], but we would run out of...resources very quickly when we need it for other things" (Sanders et al., 2011, p. 6)

A tool can therefore lead to 'inappropriate' decisions for a given organisation, such as recommending treatment with limited availability and causing unmanageable waiting times, or diverting resources away from Key Performance targets. This is more likely when PCDSS models are too prescriptive in their recommendations ('do *x*', rather than 'consider *x*, *y*, or *z*') or are based on guidelines alone and do not account for the context they are deployed in, as one decision can be appropriate from a research point of view but improper in practice.

3) **PCDSS' understanding of risk factors is limited** (Benbenishty & Treistman, 1998; Colombet et al., 2003; Jenssen et al., 2016)

   There is less evidence on risk factors or contraindications for treatment, so these are often missing in a PCDSS model. This increases the risk of a given individual receiving inappropriate treatment. Recommendations can also be risky on a contextual level. Nagpaul (2001) cites an example of a client suspected of being a victim of elder abuse: in Ohio, it is only defined as abuse if the victim is considered disabled, and is otherwise classed as maltreatment. Out of Ohio, the disability criterion does not apply. The same client would thus be classified as experiencing maltreatment or abuse depending on their location, with different resulting recommendations from a PCDSS. Regional laws also differ on whether abuse can be disclosed to a third party without consent, meaning a tool that 'did not know where it was' could potentially recommend action that was illegal.

4) **Identifying a need and specifying action for resolution creates a moral obligation to address that need** (Hunter et al., 2016; Nagpaul, 2001)

   Friction occurs when client needs are outside the immediate remit of a service. Take a PCDSS that assesses the likelihood of risk for depression and certain cancers. A mental health worker uses this tool and identifies a high risk for cervical cancer in an otherwise healthy client. The clinician has no professional mandate to recommend investigations for cervical cancer, no training to counsel the client, and may not be resourced to make a referral, yet there is a clear risk of preventable harm. This causes ethical and professional dilemmas, which are more likely in multi-disciplinary PCDSSs.

5) **"Availability of good tools alone does not ensure good craftsmanship or clinical judgment"** (Nagpaul, 2001, p. 60; Wilkinson & Himstedt, 2008)

   Nagpaul argues PCDSSs rely on clinicians' skills in gathering information and applying the tool appropriately. For instance, the Suspected Abuse Tool lists "unexplained decreases in bank account" as a marker of financial abuse, which should prompt a specific intervention. However, it is up to the skill of the clinician to elicit enough data to decide whether a given instance is indeed 'unexplained' to the extent it represents abuse. Two practitioners could thus come to different conclusions under similar circumstances.

   It is also up to the clinician to use their tools appropriately. Wilkinson & Himstedt (2008) found practitioners often ignored a web-based PCDSS and its resources altogether, and when used would sometimes engage in the "inappropriate selection of resources for use or misuse" with clients, against the recommendations of the PCDSS.

6) **The trade-off between values and validity: rigidity reduces bias, flexibility increases ownership** (Chorpita et al., 2007; Hunter et al., 2016; Nagpaul, 2001; Sanders et al., 2011; van Vliet et al., 2015)

Cultural rules are rarely explicitly built into a PCDSS, despite being a factor in healthcare outcomes (Caulfield, 2012; Lie, Lee-Rey, Gomez, Bereknyei, & Braddock, 2010; Ruiz, Hamann, Garcia, & Lee, 2015). Cultural insensitivity or bias is more likely in diverse populations, whether based on age, ethnicity, gender, diagnosis, etc. This this can lead to recommendations at odds with particular cultures.

The importance of cultural sensitivity can apply equally to clients, practitioners, and organisations. Hunter et al. (2016) cite an example of a PCDSS failing to be adopted because it did not, or was seen to be unable to, take into account the underlying values of the host organisation. In Sanders (2011) practitioners were unwilling to refer to a particular therapy—despite potential benefit to clients—as their colleagues were perceived to be only interested in 'clear-cut' cases, and they otherwise risked damaging professional relations.

Nagpaul (2008) argues using a PCDSS requires the practitioner to be aware of their clients' and their own values, beliefs, and culture, and how this will impact the assessment and intervention process, as the PCDSS will not be able to do this itself. Hunter et al. (2016) suggests these opportunities for deliberation should be built in to a tool to avoid omission of various nuances and complexities, and several studies used expert opinion and local knowledge in their algorithms (e.g. Chorpita et al., 2007; Hunter et al., 2016; van Vliet, et al., 2015). This argues "it is about setting values rather than rules" (Hunter et al., 2016, p. 583). However, this approach can also reduce the validity of a tool. Barnett, dosReis, & Riddle (2002) noted their clinicians did not take diagnoses into account when determining an intervention for aggression, which limited the amount of evidence available for integration and reduced specificity. Expert opinion is also rated lower (grade D) than research evidence such as systematic reviews (A) and cohort studies (B), according to the GRADE framework of evidence quality (Guyatt et al., 2009).

7) **Difficulty accounting for complexity** (Barnett et al., 2002; Benbenishty & Treistman, 1998; Cooley et al., 2015; Sanders et al., 2011)

   PCDSSs are less able to process exceptional, complex, or co-morbid cases as there is less research data to build valid decisional models. Recommendations can be potentially harmful if certain issues are not accounted for, such as referring someone to a physiotherapist to manage chronic pain when the client also has significant psycho-social problems. First, the tool may not take into account who could best manage a mental health issue, and second, the physiotherapist may be 'over-burdened' by a complex referral. These concerns would affect a treatment decision made by a practitioner, but may not be factored into a PCDSS.

4.5.2. Risks to PCDSS uptake

1) **Poorly understood or intimidating questions** (Clarke et al., 2010; Colombet et al., 2003; Kennedy et al., 2004)

   Questions open to interpretation are more likely to be misunderstood by the clinician or service user. This is particularly relevant for psychological factors, such as describing affective states. Personal questions can also be seen by the clinician as intimidating to ask, particularly if they are not in their traditional areas of expertise. This could include screening for mental health symptoms in non-mental health settings, or checking relevant medical factors in psychological clinics. Both intimidating and ambiguous questions are more likely to be ignored by clinicians altogether.

2) **Recommendation is not precise enough** (Benbenishty & Treistman, 1998)

   Human-based decision-making involves more than a simple binary output, even when the judgment itself is between ‘yes’ or ‘no’. Clinicians may demand the same detail from PCDSSs that would be available to them normally: how certain is it this recommendation is correct? If ‘no’, then what? Which variables are more strongly weighted? Trust in recommendations is more difficult without this transparency, so practitioners are more likely to reject the tool.

3) **System compatibility with existing routines and hardware** (Colombet et al., 2003; Huijbregts et al., 2013; Olfson et al., 2003; Sanders et al., 2011)

   This applies particularly to time, as many practitioners feel constricted by existing time pressures in their practice: “When you’ve got just 10 minutes even the seconds [are] really important” (Sanders et al., 2011, p. 6). Where PCDSSs are seen as taking time away from important activities, uptake is likely to be low.

   Altering routines to accommodate a new tool is difficult, particularly if it is used relatively infrequently, such as using a violence risk PCDSS for people with a forensic history in general practice. Irregularity encourages forgetfulness of how/to use a tool. Clinicians noted any PCDSS would be easier to use if incorporated into their existing electronic record systems, which may facilitate uptake.

4) **Clinicians trust themselves to make decisions more than they trust PCDSSs** (Benbenishty & Treistman, 1998; Colombet et al., 2003; Nagpaul, 2001)

   Clinicians tend to believe they make better decisions than algorithms, despite evidence to the contrary, making them less likely to use a PCDSS or comply with recommendations. One clinician commented “there are just too many factors that go into the decision by a human being that the [PCDSS] cannot possibly cover them" and the tool decision is "worthless" (Benbenishty & Treistman, 1998, p. 201). Benbenishty also found two thirds of clinicians "showed an overwhelming reliance on themselves as decision makers and were almost insulted at the prospect of consulting a computer for support" (p. 201), and half would not reconsider their decision if it differed from the tool. Yet the tool in question agreed with professionals more than they agreed with each other.

5) **Insufficient time for training** (Colombet et al., 2003; Nagpaul, 2001; Sanders et al., 2011)

   Clinicians who were unfamiliar with the PCDSS were less likely to use it. This also worked as a function over time, where occasional use was associated with reducing competence and overall poorer adoption.

6) **Justifying the PCDSS to service users** (Buckingham et al., 2015)

   Clients may be unwilling to engage with a PCDSS, such as completing online questionnaires, without a clear rationale. Clinicians may be asked to provide such justification, further increasing time pressures in consultations

8) **Ergonomic issues with use, e.g. navigation, speed, and intuitiveness** (Benbenishty & Treistman, 1998; Buckingham et al., 2015; Colombet et al., 2003)

   Tools that are not 'user-friendly' are harder to use, and therefore more likely to be rejected. Design issues can impact both service users and clinicians if the former needs to interact with the system; Buckingham et al. (2015) noted some clients needed practitioners to assist them with the PCDSS, meaning their usefulness could partly depend on the availability of clinicians. This would offset potential gains from the tool.

9) **What is accepted as important evidence differs** (Hunter et al., 2016; Olfson et al., 2003)

Tools can be rejected for not including sources of information considered relevant by clinicians, whether or not this is backed by research. Olfson (2003) for example found clinicians were more likely to ignore PCDSS screen results for substance abuse because these were based on client self-report measures. These had low perceived validity due to concern over 'patient denial'.

10) **'Considered useful' is not the same as 'will be used'** (Colombet et al., 2003; Hunter et al., 2016; Kennedy et al., 2004; Olfson et al., 2003; Sanders et al., 2011; Sharifi et al., 2014)

Many studies received feedback from clinicians that overall the system was thought to be valuable; however, very few guidelines were followed during testing. Sharifi et al. (2014) found 50% of clinicians felt the PCDSS increased their skill screening for tobacco use, yet found this did not affect overall screening rates. Similarly, Olfson (2003) discovered the majority of clinicians thought the tool helped them recognise clinical depression, even though their rate of detection was unchanged. (Hunter et al., 2016) found that although participants "were very positive" about their experience, "there did not appear to be significant changes arising from the application of the tools" (p. 583).

Conversely, staff in Clarke et al. (2010) believed their judgement of inpatient violence was unchanged with a PCDSS, although the study showed a decrease in the use of seclusion to manage violence whenever the PCDSS was trialled. This suggests the tool produced an effect that staff were unaware of and so could not incorporate into their assessment of its usefulness. These instances strongly argue perceptions of usefulness should not be relied upon as an indicator of uptake or absolute utility in future research.

11) **Implementation lacks (senior) support** (Benbenishty & Treistman, 1998)

    Lack of managerial commitment to implementation usually means a tool will have “died immediately” after the conclusion of a study, regardless of inherent utility (p. 202). However, “interest by the upper echelons is a necessary but insufficient condition for the instatement of a [PCDSS],” as general clinician acceptability is also key (p. 202).

12) **Making decisions about how decisions are made** (Barnett et al., 2002; Cooley et al., 2013; Hunter et al., 2016)

    Many studies found decisions were made very differently in practice versus research (indeed this is the rationale for PCDSSs), and often varied with locale. If this disagreement between tool and clinician is not settled, the PCDSS is more likely to be seen as lacking appropriate utility and rejected. However, it is unlikely to be resolved in every instance. Barnett, dosReis and Riddle (2002) found organisations varied in how they applied restrictive techniques to violent offenders, representing diverse opinions between themselves as well as the literature on safety and impact. Research by Barnett (2002) and Hunter (2016) found even priorities between stakeholders conflicted, including within the same organisation. The former study attempted to solve this through extensive collaborative discussions, leading to a single tool; the latter devised individualised versions of the same tool for each group. However, each decision will affect how the tool functions and potential buy-in, and highlights how some will always disagree with PCDSS decisions, regardless of the quality of evidence. “Sometimes there are no good solutions and the choice to be made is to live with a lesser evil” (Nagpaul, 2001, p. 78).

13) **Putting clinicians out of a job** (Wilkinson & Himstedt, 2008)

    Some professionals were opposed to particular PCDSSs that overlapped too much with their job roles. It is uncertain whether this is because they worried a PCDSS would be a poor substitute, which fits with previous risks on lack of validity, or because of job protectionism. If the latter, this creates an interesting tension between a PCDSS goal of reducing pressure on specialist services on one hand, and a clinician goal of protecting professional interests on the other.

4.5.3. Summary of risks

Risks from PCDSSs depended most on the quality and availability of evidence, while risks to uptake came more from practitioners’ reluctance to use the system. In both cases the expertise of the clinician is an important moderator, as Nagpaul (2001) points out: “These tools are guides and are only as good as the practitioner's willingness and ability to use them” (p. 74). The professional environment influenced how the tool was perceived and utilised, as “participants’ practices (and thinking) were heavily shaped by, and embedded in, complex organisational, political and relational contexts” (Hunter et al., 2016, p. 585). Tools that did not take into account these contexts were more likely to be rejected outright, and were more likely to pose risks to the organisation through inappropriate clinical recommendations.

The high likelihood of one or more of these factors being present in the deployment of any PCDSS makes it important to understand the conditions that reduce these risks. This is explored further in the next section.

4.6. Realist synthesis: conditions of PCDSS effectiveness

The synthesised findings are set out in terms of theories, themes, chains of inference, and hypotheses. For brevity this paper focuses on reporting narratives of the final hypotheses, while the rest of the analysis can be found as appendices. This is summarised below in Figure 6.

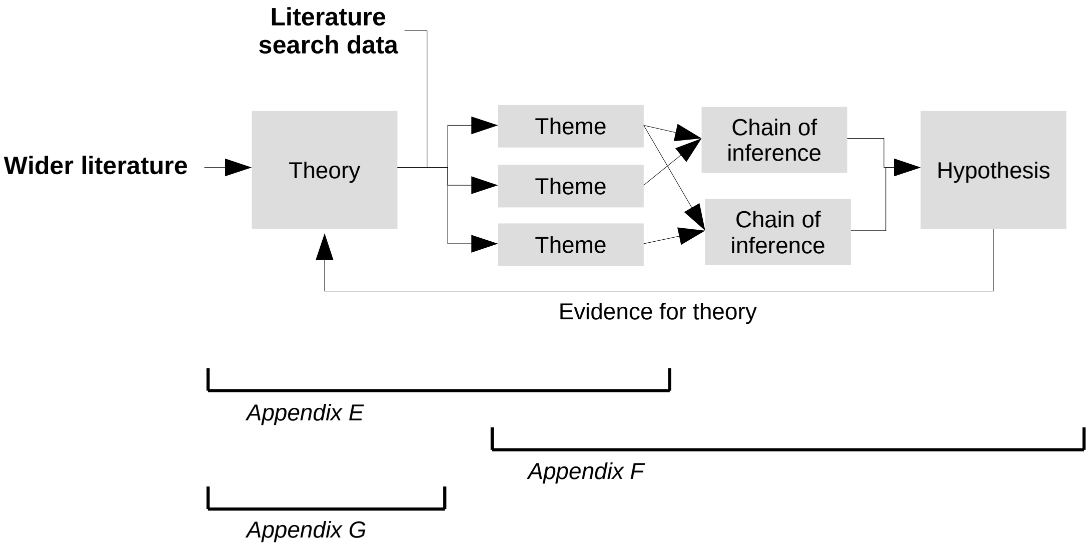


Figure 6: Flow diagram of the synthesising process and where outputs can be found in the appendices

Theories are the underlying assumptions as to why a PCDSS is expected to be effective; 12 theories were identified at the beginning of the synthesis through a process of deductive reasoning from the wider literature. Data from the previous analyses were then used to interrogate these theories, which identified themes for each theory area. Themes were next combined to make explanatory connections across studies, called chains of inference. Chains of inference are subgroupings of themes and each represents a particular connection that can be made between various evidence sources. For example, themes on 'service pressures', 'PCDSS savings' and 'competition' could be combined with knowledge from the literature to form the inference ‘success is partly resource-driven’. Chains were linked to formulate eight hypotheses on the conditions under which a PCDSS is or is not effective, which are reported in detail next.

4.6.1. Hypothesis [1]: Involvement of stakeholders early in the PCDSS design process is important to improve the chance of PCDSS use

Uptake of PCDSSs into clinical practice was often hampered by factors beyond their ability to make sound evidence-based decisions. PCDSSs need to be successful on a number of fronts in order to improve chances of being used routinely, including:

i. incorporating the values, priorities, and existing procedures of the organisation in which it is deployed into its model of decision-making. Consider referral criteria for relevant local services, waiting times for treatment, and giving precedence to Key Performance areas.

ii. integrating evidence sources valued by stakeholders—which is usually the expertise of the stakeholders themselves. As a minimum this should include managers and clinicians (to enhance uptake), ideally includes service users (to improve acceptability of recommendations), and could also extend to commissioners and specialist interest groups. This is in addition to research evidence.

iii. being ‘user-friendly’. Clinicians (and clients) are rarely motivated to invest significant time or energy to use an optional system, and the less cognitive effort required to complete a task the better the uptake. Time is a particular barrier to adoption: the longer it takes to learn a system the higher the chance it is dropped beforehand.

iv. feeling ‘owned’ by stakeholders. This enhances a sense of trust in the tool and its decisions, and that the PCDSS is more ‘like me’. It is important for key stakeholders to feel invested in the tool in order to improve adherence, prioritise resources for its use, and continue development as needed.

Each area is better addressed in the earliest stage of the tool design processes, and periodically revisited after development. Studies where these factors were only looked at after the initial research-based model had been established had more difficulty integrating them. Few studies addressed more than two areas and only one addressed all; uptake was generally poor in all but this single study. Involving stakeholders directly in the design is the easiest way to target these four factors:

> “When the potential users are part of the design team or when they feel at least involved in the development of the system, their willingness to use the resulting [PC]DSS increases manifold.”
>
> (Benbenishty & Treistman, 1998, p. 203)

This can be done using established practices from user-centred design, such as stakeholder workshops, usability testing, and good communication processes (Sharon, 2012).

4.6.2. Hypothesis [2]: PCDSSs improve outcomes for services, clients, and clinicians

In hypothesis [1], the design of a PCDSS needed to take into account the people who use it, the context in which it is used, and those it is intended to treat. The tool will, correspondingly, affect these stakeholder groups, and the following points should be considered as part of the design and evaluation process.

PCDSSs are most effective for services when they efficiently use limited resources. Several tools would make recommendations beneficial to clients, but that would deplete service resources overall; ultimately such decisions make the tool untenable. However, when care can be matched to particular client circumstances and problems, it is more likely for better care to be provided for the same investment as treatment as usual. This argues PCDSSs should prioritise conditional models of treatment ('under circumstances *x*, treat with *y*, else *z*'), such as stepped care and personalised treatment approaches. Where possible, contraindications for treatment should also be incorporated.

PCDSSs can also benefit services when they highlight additional gaps in need, as long as these can be addressed as efficiently. For instance, a decisional tool for junior doctors to assess bio-psycho-social discharge factors on a hospital ward might identify a need for linked social workers to have a PCDSS based on medical, emotional, and housing factors. This would decrease hospital readmittance rates more than either tool alone. Suitably incorporating other teams' requirements into the PCDSS model will also decrease friction at points of contact between services, for example by increasing the number of appropriate—and therefore accepted—referrals, thus improving inter-agency functioning.

Providing a more effective service overall is of obvious benefit to service users, who profit most from PCDSSs when they are more quickly allocated the care that most effectively addresses their problem of concern. Faster identification occurs when tools can identify symptoms through client feedback, e.g. via a questionnaire, that otherwise the practitioner would not be aware of (for instance asking about symptoms of ADHD is not routine in most primary care centres, so is generally identified only when problems advance and become obvious). However, identification will only be effective if the client wishes to address the problem: a referral for depression is more likely to be a waste of clinician time if the service user does not consider it an issue to be treated by that service. Motivation for treatment is enhanced when the PCDSS is matched to a context that makes sense to the client, such as a paediatrician suggesting the client stop smoking because it could harm their child, rather than a psychologist highlighting links to ill health. Early, effective treatment increases chances of recovery from any illness, so clients are more likely to recover from a mental illness with a PCDSS under these conditions.

Finally, clinicians themselves can benefit from using a PCDSS when it causes them to reflect on their decision-making. As a minimum, this occurs when the clinician sees a tool's recommendation, suggesting PCDSSs should always produce a visible decision, regardless of clinician input (or indeed whether that decision is accurate). Where the tool's decisions are based on established guidelines, this increases the chances of clinicians making guideline-based treatment decisions, although they may not be aware of a change in their behaviour. Ideally the decisional model of the PCDSS should be transparent, as this will improve clinician understanding of their own decisional process, as well as enhance their ability to explain it to clients.

4.6.3. Hypothesis [3]: Impact on mental health outcomes depends on clinician behaviour, organisational support, and evidential integration

Most PCDSSs are constructed on the basis they affect mental health outcomes primarily through making more evidence-based decisions than a clinician. This is incorrect. PCDSSs impact mental health outcomes by altering clinician behaviour, and no amount of GRADE A research will make a difference without this. Most PCDSSs that fail to improve outcomes—despite being evidence-backed—do so in the first instance because clinicians ignore it. This can be for a variety of factors including poor usability or low trust, which are expanded on elsewhere. Behaviour change can be facilitated when clinicians view PCDSSs as valuable tools they work with (rather than being imposed), which is discussed further in hypotheses [6] and [7].

The second biggest facilitator of impact is support from the organisation in which the PCDSS is used, usually by senior management. They are important in prioritising resources for the (continued) use of the new system, including clinician time. This also needs to be communicated to clinicians: in one study clinicians refused to follow tool recommendations as they believed the necessary treatment resources were not available, even though in fact extra was purchased specifically for the study. This again highlights the primary importance of clinician behaviour. Organisational support is also helpful to drum-up motivation to develop the tool over time as new research, procedures, and needs arise. Prolonged use and continuous development are unlikely to occur without assistance from management.

This support is useful in contributing to the evidence base informing the decisional model. Evidence must come from a range of sources, as described in hypotheses [1] and [7], including managerial priorities. Several studies found considerable resistance to their tools when local, organisational, and professional evidence was not built into models. This not only increases the perception of poor value, but reduces actual effectiveness by being poorly adapted to its environment. As the research context is very different to the context of clinical practice, it is important to properly incorporate a broad range of sources:

> [In research] investigators commonly test the influence of particular therapeutic practices on clinical outcomes in a highly optimized context. Such conditions may be difficult to replicate in service organizations, and the notion that all evidence-based practices are robust to these changes in context seems unlikely
>
> (Chorpita et al., 2007, p. 115)

Models based only on research evidence are therefore less likely to have an impact on outcomes.

4.6.4. Hypothesis [4]: PCDSSs are more likely to be used and adhered to when they are trusted as decision makers

Professionals quite naturally trust themselves to make professional decisions. Their decisions are largely adequate, and so intervention is, by default, regarded as unnecessary at best or insulting at worst. Any knowledge of the superiority of algorithms is generally not considered to apply to themselves. Familiarity with PCDSSs or research evidence on their efficacy is therefore not sufficient to engender use; clinicians must trust the PCDSS's decisions as they trust themselves, otherwise why use it in the first place?

The following key questions must be adequately addressed in the mind of the clinician before they will trust a PCDSS enough to use it: how well does it support decisions for complex clients; how well does it take contextual factors (organisational policy, resource availability, targets, etc.) into account; and does it do both of these better than myself [the clinician]? Critically, the PCDSS does not have to be perfect in addressing these issues, but can be imperfect and allow for correction by the clinician. Allowing individual discretion increases trust in the tool because it now includes input from a trusted source: the clinician. Conditions of trust are expanded on in hypothesis [6].

4.6.5. Hypothesis [5]: It is more important to make valued decisions than decisions that are right according to the research base

Making the 'right' decision in an absolute sense requires adopting a single point of view to the exclusion of all others. 'Right' in mental health care is thus inherently conflictual, given the number of stakeholders involved. For instance treating a person's anxiety disorder may be right for that client, but may not be right for another when it diverts resources from someone more vulnerable. Meeting targets for smoking referrals may benefit a service, but reduces the time clinicians have to assess for domestic violence. Guidelines can only recommend interventions established by research, but not experimental and potentially more beneficial treatments. Making a decision based only on one point of view, usually the research base, is common practice but more likely to be 'wrong' for a given context. Thus the different 'rights' need to be weighed up and compromises made, otherwise the tool will most likely be ineffective and rejected. Clinicians will often refuse to use a PCDSS that does not include organisational values, even though the research is sound.

> "Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful."
>
> (Box & Draper, 1987, p. 74)

Decisional models should therefore focus on maximising value to the various key stakeholders (client, clinician, and organisation), which includes but is not exclusively based on research. What is valuable can be determined through suitable stakeholder involvement (see hypothesis [1]). Flow of resources is a common area of value, and is readily addressed by most PCDSSs. With the proper design, algorithms can effectively identify early opportunities for treatment, especially when supported by the context (see hypothesis [2]), and recommend interventions that are cost-effective. Improving use of existing resources tends to bring more value to more stakeholders than increasing consumption by strict adherence to guidelines. Achieving this also helps reduce barriers to PCDSS adoption, as use of any tool entails a necessary increase in time away from other routines.

4.6.6. Hypothesis [6]: Trust in a PCDSS is related to risk, information, transparency, discretion, and personalisation

Clinicians are held responsible for the treatment of their clients, and organisations are responsible for their clinicians. Decisions on appropriate interventions thereby come with a certain amount of risk to all three parties, and there must be trust in the expertise of the decider in order to accept the recommendation. When the decider is an algorithm, the usual cues for determining trustworthiness in a human (facial expression, professional membership, etc.) are often missing, so establishing trust is difficult. When the alternative to the PCDSS is practice-as-usual, which already has a track record of being ‘good enough’, the less risky option is the more certain one, and the PCDSS is dropped. Decreasing perception of risk to improve trust can come from reallocating responsibility for decisions away from the clinician (people are more likely to take risks if they are not held responsible for poor outcomes) and by communicating good outcomes. Where responsibility is not clearly defined, practitioners will assume it lies with themselves and are more likely to disregard tool recommendations.

PCDSSs are seen as more trustworthy when they incorporate valued sources of evidence. This is not necessarily that which is valued from a research perspective (although these should be included to enhance decision effectiveness), but can also include local- and client-based sources.

> “[PC]DSS advice must be understood intuitively by decision makers, with trust in its provenance being an important factor in system adoption”.
>
> (Buckingham et al., 2015, p. 1190)

Overall, including stakeholder expertise can enhance perceptions of trust, as the tool by its nature becomes more ‘like me’. This evidence and the way it is used in decisions should be as transparent as possible, especially to clinicians, as this reduces uncertainty (which is anathema to trust). Including degrees of uncertainty and risk factors for treatment is helpful. However, if the model output is too complex, the clinician will generally not try and understand it and transparency will be effectively lost.

Feeling in control is important to enhance trust and reduce perception of risk. Practitioners are more likely use a PCDSS if they can adapt its recommendations, although this means more deviations from guidelines. Having decisions imposed on them by a tool or anyone else, especially without transparency, decreases perceptions of control, and clinicians will naturally try and restore the balance by ignoring the PCDSS where possible.

> This feels like (and is) being told to go places to perform activities as directed, not knowing why, not being part of the decision-making process, and having to report back as required. Not surprisingly, he resents it as an insult rather than an aid to intelligence...Design should be in the process, and that means putting intelligence where the work is done at the front line, not...remote from the work
>
> (Seddon, 2008, p. 129)

Such discretion also helps personalise decisions to the client and context, where PCDSSs are generally at a disadvantage. PCDSSs should demonstrate personalisation where they are able, and allow practitioners to weigh in where they are not.

4.6.7. Hypothesis [7]: PCDSSs are most effective at improving mental health outcomes when they function as part of a mutually symbiotic relationship

As demonstrated earlier, PCDSSs are not effective based on their integration of research alone (see hypotheses [3] and [5]). They require the support of clinicians and the organisation they are deployed in to be effective on any level, yet work at their best when able to draw on expertise from multiple sources. All this expertise does not have to be built directly into the algorithm; indeed this would make the model more complicated, harder to understand, and appear less trustworthy as a result. It is often enough to provide opportunities in the design of the tool for input from clinicians, clients, etc. so they can correct any errors the tool makes.

> “Rules stop the system absorbing variety”
>
> (Seddon, 2008, p. 123)

PCDSSs are more likely to make mistakes in unusual or complex cases (as there is less evidence available on effective care), where human decisions are more valuable. Humans on the other hand make more mistakes when predicting outcomes in general or have been qualified for a period of time ('expert' thereby being somewhat of a paradox). Integration of expert, contextual, and research knowledge should still be incorporated into the PCDSS where possible, but not at the expense of transparency, comprehension, or opportunities for discretion.

4.6.8. Hypothesis [8]: PCDSSs are more likely to improve mental health outcomes when they are matched to specific contexts and problems

When PCDSSs are used generally to assess for specific problems, they will more often identify people for whom treatment is inappropriate. For instance, few people attending a primary care clinic will have symptoms of and desire treatment for Obsessive Compulsive Disorder, while more people attending a mental health centre will. Being screened for treatment can even be intrusive in contexts where it is not expected, such as for post-traumatic stress in a diabetes clinic. PCDSS usefulness can thus be increased when it is matched to the setting.

PCDSSs can be more effective when they target specific problems, such as feeling low or fatigued, rather than operating at the diagnostic level, such as depression. Where such modular approaches are possible, this wastes fewer resources on treating problems that are not present (not everyone with post-traumatic stress has nightmares for example), and can lead to improvement more quickly by targeting symptoms of concern.

4.7. Summary of results

Studies show mixed efficacy for PCDSSs to improve mental health outcomes. No clear pattern of results was established from a narrative study of experimental research, although a meta-analysis suggested PCDSSs are effective for improving counselling rates but not assessments for smoking. An examination of clinician adherence to tool recommendations indicated PCDSSs are not feasible as they are currently deployed, with only one study meeting criteria for both assessment and intervention adherence. This linked to risks identified from PCDSSs and to their uptake in general practice, as many studies reported poor adherence due to concerns over their ability to appropriately model for complexity and context.

A realist synthesis was used to expand upon the conditions for these results, suggesting feasibility and risks could be improved with stakeholder-driven design processes, incorporation of a variety of evidence sources, and a context-driven approach to deployment.

> Although algorithms have many advantages, such as decreasing the variability in practice and increasing the ease of evaluation, the disadvantages are the limited number of variables that can be addressed, the lack of consensus among clinicians, and the inability to generalize to populations with co-morbid disorders or whose symptoms do not meet specific diagnostic criteria
>
> (Barnett, dosReis, & Riddle, 2002, p. 899)

PCDSSs were most effective when adapted to their organisational context and enable the clinician to exercise their own expertise. They were least effective in practice when overly prescriptive and based only on research data.

5. Discussion

> "As we have seen, interventions never run smoothly. They are subject to unforeseen consequences as a result of resistance, negotiation, adaptation, borrowing, feedback and, above all, context, context, context."
>
> (Pawson, Greenhalgh, Harvey, & Walshe, 2004, p. 16)

This study began assuming PCDSSs could be successful due to a research-driven, 'top-down' approach: the tools generate statistically superior decisions based on experimental studies, and these are followed faithfully by the clinician. However, the evidence gathered argues this is rarely the case, and not sufficient for long-term efficacy. Rather it is suggested PCDSSs function by working collaboratively with the clinician, and appropriately integrating research, expert and contextual evidence to form the best 'ecological fit' between tool, practitioner, organisation, and client. This approach encourages clinicians to reflect on their decision-making processes, which increases adherence to research-based evidence while also allowing for professional discretion in cases where research is less strong. Under these circumstances PCDSSs can improve outcomes for: clients by matching problems of concern to the most effective available intervention, leading to a better treatment response; clinicians by creating opportunities to bring their judgements closer to best practice; and services by making best use of resources, improving collaboration between services, and highlighting additional decisional needs. Efficacy is thus intrinsically bound to context, as Pawson states:

> "The success of an intervention is not simply a question of the merit of its underlying ideas but depends, of course, on the individuals, interpersonal relationships, institutions and infrastructures through which and in which the intervention is delivered."
>
> (Pawson et al., 2004, p. iii)

5.1. Strengths and limitations of the review

The review gives a broad overview of the efficacy of PCDSSs in mental health outcomes and an in-depth interrogation of the available data. This approach generated a new understanding of decision support systems and highlights practical ways that could improve future PCDSSs, regardless of the psychological issue under consideration. This generalisability has been at the expense of specificity, as useful conclusions on the efficacy of PCDSSs for particular outcomes or disorders—with the potential exception of smoking—was not possible with the identified studies due to heterogeneity.

This study was conducted with a reasonable degree of methodological rigour, such as following a protocol-driven search (Higgins & Green, 2011) and systematic synthesis (Pawson et al., 2004; Rycroft-Malone et al., 2012). The analysis could have been improved with the validation of codes by a second qualitative researcher. The impact of this has been reduced through the 'expert consultation process' and a transparent analysis process from paper to conclusions. However, some caution should be exercised regarding the method as it is a 'hybrid' of both pre-defined and iterative processes. These are normally regarded as separate strategies (Pawson et al., 2004) and it is not usual to combine them. This approach was chosen to better reduce bias and usefully analyse studies with a large degree of heterogeneity, at the risk of methodological consistency. Had a single approach been used, either the findings on efficacy would have been limited to a narrative description, or the evidence sources used would have included shifted the focus from efficacy to policy.

The review was unable to comment on the general sensitivity or specificity of algorithms in detecting mental health symptoms, as this was rarely reported. Their accuracy was often compared to clinician ratings (which are themselves open to bias), so PCDDS accuracy was normally phrased in terms of the number of cases that agreed with clinicians. The degree to which low sensitivity affects different PCDSSs would be uneven, as algorithms with few data inputs would tend to create larger errors. On the other hand, poor specificity is more likely to threaten clinician use of PCDSSs, as their recommendations would be seen as less trustworthy and more likely to waste valuable resources. This would therefore be a useful area to follow up in the future.

The evidence found suggested a range of conditions on the efficacy of PCDSSs regarding clinicians and services. However, very little data were reported from service users, such as their views on algorithms. This argues the results of this study are biased in favour of interpretations by researchers and practitioners, and should be used carefully with patients. Similarly, many of the clinicians themselves were medical practitioners and not psychologists—although they performed psychological work—which may skew the findings in favour of medical models. However, the preponderance of tools built for physical health settings could be argued to help with parity of esteem by encouraging clinicians of both psychological and medical disciplines to think more equitably about physical and mental health:

> Parity of esteem is...making sure that we are just as focused on improving mental as physical health and that patients with mental health problems don't suffer inequalities, either because of the mental health problem itself or because they then

> don't get the best care for their physical health problems.
>
> (NHS England, 2013, p. 8)

Therefore, parity of esteem should work both ways, and the applicability of findings from this research to mental and physical health settings will hopefully become an advantage rather than a limitation in the future.

5.2. Future research

Currently little is reported about the rate of errors in PCDSSs or who is responsible for them. These are potential barriers to the adoption of PCDSSs more generally, and future research on the effect of varying accountability on both uptake and long-term efficacy would be interesting. Openly assigning responsibility to clinicians, organisations, commissioners, or clients may alter viability in different ways, especially as all these groups are stakeholders in the research. The impact of target area on adoption and accountability could also generate intriguing findings; no study for instance discussed how the practical significance of a low error rate on identifying low mood might compare to the significance of the same error rate for identifying risk of violence to others. The potential costs of mistakes are different, which may impact the use or misuse of tools.

The search for papers revealed a huge number of algorithms currently available but without evidence of an impact on patient outcomes. A significant part of future research can thus be the testing of existing tools without necessarily developing new ones, although this paper still argues adaptation to local standards is important. However, adjusting PCDSSs to context raises a potential confounder for existing research methodologies: how can efficacy be experimentally established if context cannot be controlled for? Five sites could use the same basic tool, but have five different model versions. Even if usefulness was verified for all of them, the same efficacy would not necessarily apply to site number six. It would be of some gain to establish a benchmark method for assessing algorithms so deployed in order to generate meaningful results without endless testing.

One last direction for research could be the comparison between mono- and multi-disciplinary PCDSSs on mental health outcomes. Evidence in this paper suggests a multi-disciplinary tool is more effective when it identifies more issues of concern in a timely manner and improves interactions between and within teams.

> Some past [P]CDSS approaches have been specific to mental health issues and, as a result, may have been intrusive and disruptive to the usual processes of care...Using a holistic [P]CDSS holds much promise for introducing better evidence-based care and ongoing chronic care management
>
> (Carroll et al., 2013b, p. e628)

As with parity of esteem, mental and physical health are widely held to be mutually impactful, and treating both with one tool could save a proportion of the 45% additional cost in healthcare estimated to be caused by co-morbidities (Naylor et al., 2012). However, more complex models are difficult to understand and less informed by research, which could negatively impact overall efficacy.

5.3. Conclusion

This study began by supposing the efficacy of PCDSSs was limited to the programmable system, firmly based in the validity of its underlying algorithm. Since algorithmic supremacy to human decision-making was well established and yet uptake was poor, this was only evidence of human fallibility and possibly that creeping suspicion of clever technology that pervades every science fiction contemplation. However, the review has also highlighted the fallibility of algorithms to account for complexity, and the current necessity for them to work with their fleshy counterparts to make the most sense of the world. There are numerous risks from both man and machine, yet hope remains that suitable design approaches can facilitate the symbiotic relationship necessary for the best mental health outcomes. Thus it is the cooperation of the complexity of A, C, T, and G with the simplicity of 1 and 0 that make PCDSSs most effective.

6. References

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., … Rush, J. D. (2006). The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction. *The Counseling Psychologist, 34*(3), 341-382.

Adli, M., Bauer, M., & Rush, A. J. (2006). Algorithms and Collaborative-care Systems for Depression: Are They Effective and Why?. A Systematic Review. *Biological Psychiatry*. http://doi.org/10.1016/j.biopsych.2006.05.010

Barnett, S. R., dosReis, S., & Riddle, M. a. (2002). Improving the management of acute aggression in state residential and inpatient psychiatric facilities for youths. *Journal of the American Academy of Child and Adolescent Psychiatry*, *41*(8), 897–905.

Bastardi, A., & Shafir, E. (1998). On the pursuit and misuse of useless information. *Journal of Personality and Social Psychology, 75*(1), 19-32.

Benbenishty, R., & Treistman, R. (1998). The development and evaluation of a hybrid decision support system for clinical decision making: The case of discharge from the military. *Social Work Research*, *22*(4), 195–204.

Bowles, K. H., Hanlon, A., Holland, D., Potashnik, S. L., & Topaz, M. (2014). Impact of Discharge Planning Decision Support on Time to Readmission Among. *Professional Case Management*, *19*(1), 29–38.

Box, G. E. P., & Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*. United States: John Wiley and Sons Inc.

Buckingham, C. D., Adams, A., Vail, L., Kumar, A., Ahmed, A., Whelan, A., & Karasouli, E. (2015). Integrating service user and practitioner expertise within a web-based system for collaborative mental-health risk and safety management. *Patient Education and Counseling*, *98*(10), 1189–1196.

Bugental, D. B., & Happaney, K. (2004). Predicting Infant Maltreatment in Low-Income Families: The Interactive Effects of Maternal Attributions and Child Status at Birth. *Developmental Psychology, 40*(2), 234-243.

Burt, A., & Volchenboum, S. (2018). How Health Care Changes When Algorithms Start Making Diagnoses. Harvard Business Review. Available from http://hbr.org/

Carroll, A. E., Biondich, P., Anand, V., Dugan, T. M., & Downs, S. M. (2013a). A randomized controlled trial of screening for maternal depression with a clinical decision support system. *Journal of the American Medical Informatics Association*, *20*(2), 311–316.

Carroll, A. E., Bauer, N. S., Dugan, T. M., Anand, V., Saha, C., & Downs, S. M. (2013b). Use of a computerized decision aid for ADHD diagnosis: a randomized controlled trial. *Pediatrics*, *132*(3), e623-9.

Caulfield, K. J. (2012). *Relationships among Self-Efficacy, Health Beliefs and the Self-Management Behaviors of Healthy Eating and Physical Activity among Adults with Type 2 Diabetes.* George Mason University, ProQuest Dissertations Publishing.

Chase, J. (2014). A clinical decision algorithm for hospital inpatients with impaired decision-making capacity. *Journal of Hospital Medicine*, *9*(8), 527–532.

Chan, M. (2017). *Opening remarks at the Artificial Intelligence for Good global summit*. Geneva: World Health Organisation. Available from: http://www.who.int/dg/speeches/2017/artificial-intelligence-summit/en/

Chorpita, B. F., Bernstein, A., Daleiden, E. L., Weisz, J., Chorpita, B., Gibbons, R., … Schoenwald, S. (2007). Driving with roadmaps and dashboards: Using information resources to structure the decision models in service organizations. *Administration and Policy in Mental Health and Mental Health Services Research*, *35*(1–2), 114–123.

Clark, D. A. (1992). Human expertise, statistical models, and knowledge systems. In G. Wright & F. Bolger (Eds.), *Expertise and decision support* (p. 227-249). New York: Plenum Press.

Clarke, D. E., Brown, A. M., & Griffith, P. (2010). The Brøset Violence Checklist: Clinical utility in a secure psychiatric intensive care setting. *Journal of Psychiatric and Mental Health Nursing*, *17*(7), 614–620.

Colombet, I., Dart, T., Leneveut, L., Zunino, S., Menard, J., & Chatellier, G. (2003). A computer decision aid for medical prevention: a pilot qualitative study of the Personalized Estimate of Risks (EsPeR) system. *BMC Medical Informatics and Decision Making*, *3*(13).

Cooley, M. E., Blonquist, T. M., Catalano, P. J., Lobach, D. F., Halpenny, B., McCorkle, R., … Abrahm, J. L. (2015). Feasibility of Using Algorithm-Based Clinical Decision Support for Symptom Assessment and Management in Lung Cancer. *Journal of Pain and Symptom Management*, *49*(1), 13–26.

Cooley, M. E., Lobach, D. F., Johns, E., Halpenny, B., Saunders, T. A., Del Fiol, G., … Abrahm, J. L. (2013). Creating computable algorithms for symptom management in an outpatient thoracic oncology setting. *Journal of Pain and Symptom Management*, *46*(6), 911–924.

Foster, N. E., Mullis, R., Hill, J. C., Lewis, M., Whitehurst, D. G. T., Doyle, C., … Hay, E. M. (2014). Effect of Stratified Care for Low Back Pain in Family Practice (IMPaCT Back): A

Prospective Population-Based Sequential Comparison. *Annals of Family Medicine*, *12*(2), 102–112.

Garg, A. X., Adhikari, N. K. J., McDonald, H., Rosas-Arellano, M. P., Devereaux, P. J., Beyene, J., … Haynes, R. B. (2005). Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: A systematic review. *Journal of the American Medical Association*. http://doi.org/10.1001/jama.293.10.1223

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., Nelson, C., Snitz, B. E., & Grove, M. (2000). Clinical Versus Mechanical Prediction : A Meta - Analysis. *Psychological Assessment, 12*(1), 19-30.

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schünemann, H. J. (2009). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *Biomedical Journal, 337*(7650), 924-926.

Hannan, C., Lambert, M. J., Harmon, C., Nielsen, S. L., Smart, D. W., Shimokawa, K., & Sutton, S. W. (2005). A lab test and algorithms for identifying clients at risk for treatment failure. *Journal of Clinical Psychology*, *61*(2), 155–163.

Hemmelgarn, A. L., Glisson, C., & James, L. R. (2006). Organizational Culture and Climate: Implications for Services and Interventions Research. *Clinical Psychology: Science and Practice*, *13*(1), 73–89.

Higgins, J. P. T., & Green, S. (2011). *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0*. The Cochrane Collaboration. Available from: www.cochrane-handbook.org

Huijbregts, K. M. L., De Jong, F. J., Van Marwijk, H. W. J., Beekman, A. T. F., Adèr, H. J., Hakkaart-Van Roijen, L., … Van Der Feltz-Cornelis, C. M. (2013). A target-driven collaborative care model for Major Depressive Disorder is effective in primary care in the Netherlands. A randomized clinical trial from the depression initiative. *Journal of Affective Disorders*, *146*(3), 328–337.

Hunter, D. J., Marks, L., Brown, J., Scalabrini, S., Salway, S., Vale, L., … Payne, N. (2016). The potential value of priority-setting methods in public health investment decisions: qualitative findings from three English local authorities. *Critical Public Health*, *26*(5), 578-587.

IBM. (2015). *Deep Blue*. Available from: http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/

Iyengar, M. S. (2009). *The Medical Algorithms Project. European Spreadsheet Risks Interest Group*. Available from: https://arxiv.org/ftp/arxiv/papers/0908/0908.0932.pdf

Jenssen, B. P., Shelov, E. D., Bonafide, C. P., Bernstein, S. L., Fiks, A. G., & Bryant-Stephens, T. (2016). Clinical Decision Support Tool for Parental Tobacco Treatment in Hospitalized Children. *Applied Clinical Informatics*, *7*(2), 399–411.

Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica, 47*(2), 263-292.

Karppi, T., & Crawford, K. (2016). Social Media, Financial Algorithms and the Hack Crash. *Theory, Culture & Society, 33*(1), 73-92.

Kennedy, C., Finkelstein, N., Hutchins, E., & Mahoney, J. (2004). Improving screening for alcohol use during pregnancy: the Massachusetts ASAP program. *Maternal and Child Health Journal*, *8*(3), 137–147.

Kosover, A., Yorkin, B., Johnson, B., Sikes, C. (Producers), & Villenuve, D. (Director). (2017). *Blade Runner 2049* (Motion picture). United States: Warner Bros. Entertainment Inc.

Lahey, B. B., Loeber, R., Burke, J. D., & Applegate, B. (2005). Predicting Future Antisocial Personality Disorder in Males From a Clinical Assessment in Childhood. *Journal of Consulting and Clinical Psychology,73*(3), 389-399.

Lenfant, C. (2003). Clinical Research to Clinical Practice — Lost in Translation? *New England Journal of Medicine, 349,* 868-874.

Lie, D. A., Lee-Rey, E., Gomez, A., Bereknyei, S., & Braddock, C. H. (2010). Does cultural competency training of health professionals improve patient outcomes? A systematic review and proposed algorithm for future research. *Journal of General Internal Medicine, 26*(3), 317–325.

Lobach, D. F., Johns, E. B., Halpenny, B., Saunders, T.-A., Brzozowski, J., Del Fiol, G., … Cooley, M. E. (2016). Increasing Complexity in Rule-Based Clinical Decision Support: The Symptom Assessment and Management Intervention. *Journal of Medical Internet Research Medical Informatics, 4*(4), e36.

Lugtenberg, M., Burgers, J. S., & Westert, G. P. (2009). Effects of evidence-based clinical practice guidelines on quality of care: a systematic review. *Quality and Safety in Health Care, 18*(5), 385-392.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Minnesota: University of Minnesota Press.

Meehl, P. E. (1986). Causes and Effects of My Disturbing Little Book. *Journal of Personality Assessment, 50*(3), 370-375.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta analyses: The Prisma Statement. *PLoS Med*, *6*(6).

Nagpaul, K. (2001). Application of Elder Abuse Screening Tools and Referral Protocol: Techniques and Clinical Considerations. *Journal of Elder Abuse & Neglect*, *13*(2), 59–78.

Naylor, C., Parsonage, M., McDaid, D., Knapp, M., Fossey, M., & Galea, A. (2012). *Long-term conditions and mental health: the cost of co-morbidities*. London: The King's Fund.

NHS England. (2013). *Everyone Counts: Planning for patients 2014/15 to 2018/19*. Leeds: NHS England.

Olfson, M., Tobin, J. N., Cassells, A., & Weissman, M. (2003). Improving the detection of drug abuse, alcohol abuse, and depression in community health centers. *Journal of Health Care for the Poor and Underserved*, *14*(3), 386–402.

Pawson, R., Greenhalgh, T., Harvey, G., & Walshe, K. (2004). Realist synthesis - an introduction. *Economic and Social Research Council Research Methods Programme*, 1–46.

Pluye, P., Robert, E., Cargo, M., & Bartlett, G. (2011). *Proposal: A mixed methods appraisal tool for systematic mixed studies reviews*. Montréal: McGill University.

Public Health England. (2014). *From evidence into action: opportunities to protect and improve the nation's health*. London: PHE Publications.

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., … Dean, J. (2018). Scalable and accurate depp learning with electronic health records. *npj Digital Medicine, 1*(18).

Review Manager (RevMan). 2014. [Computer program]. *Version 5.3*. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration.

Rindal, D. B., Rush, W. A., Schleyer, T. K. L., Kirshner, M., Boyle, R. G., Thoele, M. J., … Huntley, C. L. (2013). Computer-assisted guidance for dental office tobacco-cessation counseling: A randomized controlled trial. *American Journal of Preventive Medicine*, *44*(3), 260–264.

Robb, J. (2017). *How Algorithms and Authoritarianism Created a Corporate Nightmare at United*. Global Guerrillas. Available from: http://globalguerrillas.typepad.com/globalguerrillas/2017/04/algorithmic-dystopia.html

Rollman, B. L., Hanusa, B. H., Lowe, H. J., Gilbert, T., Kapoor, W. N., & Schulberg, H. C. (2002). A randomized trial using computerized decision support to improve treatment of major depression in primary care. *Journal of General Internal Medicine*, *17*, 493–503.

Ruiz, J. M., Hamann, H. A., Garcia, J., & Lee, S. J. C. (2015). The psychology of health: Physical health and the role of culture and behavior in Mexican Americans. In Y. Caldera and E. Lindsey (Eds) *Mexican American children and families: Multidisciplinary perspectives.* London: Routledge Press.

Rycroft-Malone, J., McCormack, B., Hutchinson, A. M., DeCorby, K., Bucknall, T. K., Kent, B., … Wilson, V. (2012). Realist synthesis: illustrating the method for implementation research. *Implementation Science*, *7*(1), 33.

Sanders, T., Foster, N. E., & Ong, B. N. (2011). Perceptions of general practitioners towards the use of a new system for treating back pain: a qualitative interview study. *Biomed Central Medicine*, *9*(49).

Scheirer, M. a. (2005). Is Sustainability Possible? A Review and Commentary on Empirical Studies of Program Sustainability. *American Journal of Evaluation, 26*(3).

Seddon, J. (2008). *Systems Thinking in the Public Sector: The failure of the reform regime and a manifesto for a better way*. Charmouth: Triarchy Press Limited

Sharifi, M., Adams, W. G., Winickoff, J. P., Guo, J., Reid, M., & Boynton-Jarrett, R. (2014). Enhancing the electronic health record to increase counseling and quit-line referral for parents who smoke. *Academic Pediatrics*, *14*(5), 478–484.

Sharon, T. (2012). *It's Our Research: Getting Stakeholder Buy-in for User Experience Research Projects*. London: Morgan Kaufman.

Shortell, S. M., Bennett, C. L., & Byck, G. R. (1998). Assessing the impact of continuous quality improvement on clinical practice: What it will take to accelerate progress. *Milbank Quarterly, 76*(4), 593-624.

Simonite, T. (2017). *Machines Taught by Photos Learn a Sexist View of Women*. Wired. Available from: https://www.wired.com/story/machines-taught-by-photos-learn-a-sexist-view-of-women/

Snilstveit, B., Oliver, S., & Vojtkova, M. (2012). Narrative approaches to systematic review and synthesis of evidence for international development policy and practice. *Journal of Development Effectiveness*.

Stallvik, M., Gastfriend, D. R., & Nordahl, H. M. (2015). Matching patients with substance use disorder to optimal level of care with the ASAM Criteria software. *Journal of Substance Use*, *20*(6), 389–398.

Thomas, H. V., Watson, M., Bell, T., Lyons, I., Lloyd, K., Weich, S., … Lewis, G. (2004). Computerised patient-specific guidelines for management of common mental disorders in primary care: a randomised controlled trial. *The British Journal of General Practice: The Journal of the Royal College of General Practitioners*, *54*(508), 832.

Tolin, D. F., Diefenbach, G. J., & Gilliam, C. M. (2011). Stepped care versus standard cognitive-behavioral therapy for obsessive-compulsive disorder: A preliminary study of efficacy and costs. *Depression and Anxiety*, *28*(4), 314–323.

Tolin, D. F., Diefenbach, G. J., Maltby, N., & Hannan, S. (2005). Stepped care for obsessive-compulsive disorder: A pilot study. *Cognitive and Behavioral Practice*, *12*(4), 403–414.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*, 433-460.

Tversky, A., & Kahneman, D. (1973). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131.

van Vliet, L. M., Harding, R., Bausewein, C., Payne, S., & Higginson, I. J. (2015). How should we manage information needs, family anxiety, depression, and breathlessness for those affected by advanced disease: development of a Clinical Decision Support Tool using a Delphi design. *Biomed Central Medicine*, *13*(1), 263.

Weisz, J. R., Chorpita, B. F., Palinkas, L. A., Schoenwald, S. K., Miranda, J., Bearman, S. K., … Mayberg, S. (2012). Testing Standard and Modular Designs for Psychotherapy Treating Depression, Anxiety, and Conduct Problems in Youth: A Randomized Effectiveness Trial. *Archives of General Psychiatry*, *69*(3), 274–282.

Wilkinson, S., & Himstedt, K. (2008). Establishing an innovative model of nutrition and dietetic care for a mental health service through collaboration with non-nutrition healthcare workers. *Nutrition and Dietetics*, *65*(4), 279–283.

Woodworth, R. S. (1917). *Personal data sheet*. Chicago: C. H. Stoelting Company.

Zigmond, A., & Snaith, R. (1983). The hospital anxiety and depression scale. *Acta Psychiatrica Scandinavica, 67*(6), 361–370.

PART TWO—EMPIRICAL PAPER

USING A PATIENT PROFILING TOOL TO PREDICT AND ENHANCE THERAPY OUTCOMES: OMPARISON OF ALLOCATION METHODS

7. Abstract

Aims: The study explored the potential for improving the clinical utility of a patient profiling algorithm in a primary care mental health service (IAPT) compared to treatment allocation as usual. This included examining whether the algorithm's predictive accuracy could be enhanced by incorporating posterior probabilities—the residual chances of belonging to particular profiles—into profile calculations to reduce rates of allocation bias, and then determining if this was the most effective way to assign service users to treatment compared to standard practice.

Method: In Stage One of the analysis, three models (two profile calculations and one sub-grouping analysis) incorporating posterior probabilities were compared to the original profiling model on their ability to predict reliable recovery, improvement, and deterioration in service users based on depression and anxiety scores, and dropout from treatment. The chances of achieving these outcomes for a given treatment intensity and profile was examined using odds ratios. The model most likely to effect beneficial change in these outcomes was used to calculate the proportion of clients historically matched to their recommended treatment intensity, and Stage Two looked at how much of the change in IAPT recovery rates over time could be accounted for by this and service-level factors. In Stage Three, the models were match-controlled to allocation as usual on a more recent dataset and a one-way ANOVA used to analyse outcomes.

Results: The original algorithm and a Maximum Likelihood correction were equally effective at improving clinical outcomes based on an early dataset. A regression analysis indicated reliable recovery rates but not reliable improvement could be partially accounted for by natural improvements in matching clients to their most effective treatments, Beta=0.60, $p<0.05$, also suggesting service users were more likely to receive the treatment the algorithm would recommend for them over time. There was no significant difference in recovery or improvement rates between clients matched to algorithm-recommend treatment intensity or allocation as usual, although the former was significantly cheaper.

Conclusions: Including posterior probabilities did not improve the usefulness of the original algorithm, nor did it outperform IAPT professionals in improving therapeutic outcomes. However, its use could lead to significant savings without impacting recovery rates by recommending less expensive but equally effective treatment intensities.

8. Introduction

With almost 1,500,000 referrals in 2018 (Nuffield Trust, 2018), Improving Access to Psychological Therapies (IAPT) is the biggest provider of primary mental health care in the UK. Since rolling out nationwide in 2012, IAPT is judged by its ability to help service users recover from clinically-significant levels of depression and anxiety symptoms, and all IAPT sites are held to a 50% recovery standard (National Health Service England, 2018). Although national recovery rates have improved over time, many services still struggle to achieve this. For example in the year to 2018, 30% of Clinical Commissioning Groups with IAPT sites did not meet their 50% target (Public Health England, 2018). Yet even though the national average stood at 50.8% of service users moving into recovery (PHE, 2018), this means almost half of IAPT clients did not recover following a course of therapy. Of all clients referred, this is only 20%.

In addition to the therapy received, service-level and individual variables affect the likelihood of recovery. Service-level factors in IAPT have been examined by Clark et al. (2018), who determined factors such as social deprivation, number of treatment sessions and proportion of referrals treated are associated with chances of symptom improvement. At the individual level, research around personalised treatment has been gaining momentum, fuelled partly by the large amount of clinical data IAPT routinely collects for every client. The idea of personalised treatment recognises that no one treatment works for everyone, nor works equally well if it did. Tailoring mental health interventions at the individual level promises better therapeutic outcomes without developing new therapies, and it is only in the last few decades that both computing power and, ironically, large-scale data collection has advanced sufficiently to realistically achieve this. Using data from thousands of clients can create algorithmic protocols for the individual, and these can outperform professionals in determining what works for whom on a range of psychological issues including depression (Huijbregts et al., 2013), substance misuse (Stallvik, Gastfriend, & Nordahl, 2015), and violence (Clarke, Brown, & Griffith, 2010).

8.1. The current Predictive Clinical Decision Support System

Predictive Clinical Decision Support Systems are algorithms used to assist professionals make better decisions about psychological interventions by anticipating what will be effective in a given situation. The situation includes both the person affected by the decision and the setting in which the decision is taken. The present study investigates the support system developed by Saunders, Cape, Fearon, and Pilling (2016), which uses self-completed questionnaires on clinical symptoms and socio-demographic information to recommend the IAPT treatment intensity most likely to result in recovery or reliable change for a given service user. The questionnaire data is used to allocate clients to one of eight profiles, each of which have specific odds of benefiting from low- or high-intensity therapy (see Table 5 below).

Table 5

Latent profile descriptions and therapeutic outcomes; from Saunders et al. (2016)

Profile	Key Probable Characteristics	Therapeutic outcomes in IAPT
LP1	Low depression, low anxiety, in their 30s, impaired functioning*, no antidepressant medication, no benefits, no phobia	Most likely to recover/show reliable change with high-intensity therapy (79%), likely to benefit from therapy**
LP2	Moderate depression, moderate anxiety, in their early 30s, impaired functioning, no antidepressant medication, no benefits, no phobia	Most likely to recover/show reliable change with high-intensity therapy (66%), likely to benefit from therapy
LP3	Low depression, low anxiety, in their 60s, typical functioning, no antidepressant medication, no benefits, no phobia	Equally likely to recover/show reliable change with high- or low-intensity therapy (72%), likely to benefit from therapy
LP4	Moderate depression, moderate anxiety, in their 50s, impaired functioning, no antidepressant medication, no benefits, no phobia	Most likely to recover/show reliable change with low-intensity therapy (64%), likely to benefit from therapy
LP5	Moderate depression, severe anxiety, in their 40s, moderately-impaired functioning, prescribed antidepressant medication, no benefits, has a phobia	Most likely to recover/show reliable change with high-intensity therapy (44%), unlikely to benefit from therapy
LP6	Moderate depression, low anxiety, in their 40s, moderately-impaired functioning, prescribed antidepressant medication, in receipt of benefits, has a phobia	Most likely to recover/show reliable change with high-intensity therapy (56%), likely to benefit from therapy
LP7	Severe depression, severe anxiety, in their early 40s, severely-impaired functioning, prescribed antidepressant medication, in receipt of benefits, has a phobia	Most likely to recover/show reliable change with high-intensity therapy (24%), very unlikely to benefit from therapy
LP8	Moderate depression, severe anxiety, in their late 20s, moderately-impaired functioning, no benefits, has a phobia	Most likely to recover/show reliable change with high-intensity therapy (44%), unlikely to benefit from therapy

*Functioning, as measured by the W&SAS, does not have explicit cut-offs. However, for ease of understanding they are described here as severely-impaired (score >20), moderately-impaired (10-20), and typical (<10), following Mundt, Marks, Shear, and Greist (2002).

****Benefit from therapy is understood here as any reliable reduction in symptoms** at the end of an intervention (i.e. includes reliable recovery and reliable change), as measured by the PHQ-9 and GAD-7 on the Minimum Data Set.

The algorithm uses data routinely collected in all IAPT services, known as the Minimum Data Set (MDS; see Appendix H and the Observed Variables in Figure 7 below). Saunders' analysis of these data in a particular IAPT site suggested eight distinct underlying patterns of response, or latent profiles (LP), which cannot otherwise be directly measured. Membership of these profiles was found to differentially predict therapeutic outcomes including recovery from clinical levels of depression or anxiety symptoms, deterioration or improvement in symptoms, and dropout from therapy (see Figure 7 below). How variables are translated into profiles is important, as different methods can result in distinct therapy recommendations.

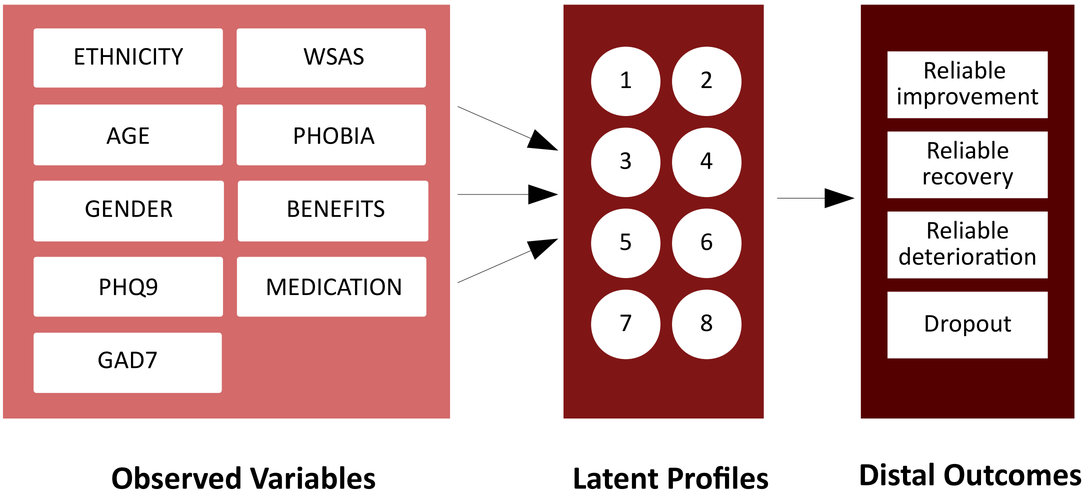


Figure 7: The relationship between variables, profiles, and outcomes in the current latent profile analysis

The algorithm takes the observed clinical and socio-demographic variables and checks them for conformity with eight previously identified patterns of response, corresponding to the eight LPs. Any specific individual from a given dataset has a probability between zero and one of belonging to each LP. Typically, the chance they belong to more than one profile in greater than zero. The likelihoods of belonging to each profile are collectively known as posterior probabilities, which are the chances of obtaining the values within a parameter (i.e. the probabilities of belonging to LP1-LP8) given an observed dataset (i.e. the MDS). These are calculated using Bayes Theorem, which describes the probability of an event based on prior knowledge of conditions that are presumed to be related (Bayes, 1763).

The most common analytic approach is to assign individuals to a single LP, requiring a decision on how to allocate to profiles (as there is a probability of membership to any of the LPs). Saunders et al. (2016) used a fixed-probability model, which takes the LP with the highest probability (also known as the ‘maximum a posteriori’, or MAP) and assigns it a *p* value of 1; allocation to this LP is now treated as a certainty in subsequent analyses (Nagin, 2005). The MAP model assigns the most likely profile membership and otherwise ignores posterior probabilities (PP).

Fixing the LP in this way introduces a certain amount of error and imprecision related to the size of the PP that are discounted. Several models attempt to take these into account when computing profile membership (referred to here as +PP classifications). Bolck, Croon, and Hagenaars (2004) for instance use a weighted function to assess the relationship between profile membership and external variables, rather than treating membership as known. Vermunt (2010) has modified this approach (mBCH) for continuous variables, and also introduced another alternative +PP, the Maximum Likelihood correction (ML), which treats LP as a variable with known error probabilities. A simplified example case is shown below[2] in Figure 8.

[2] Mathematical accuracy has been sacrificed for the sake of comparison, and the reader is urged to consult the original papers for the most correct explanation

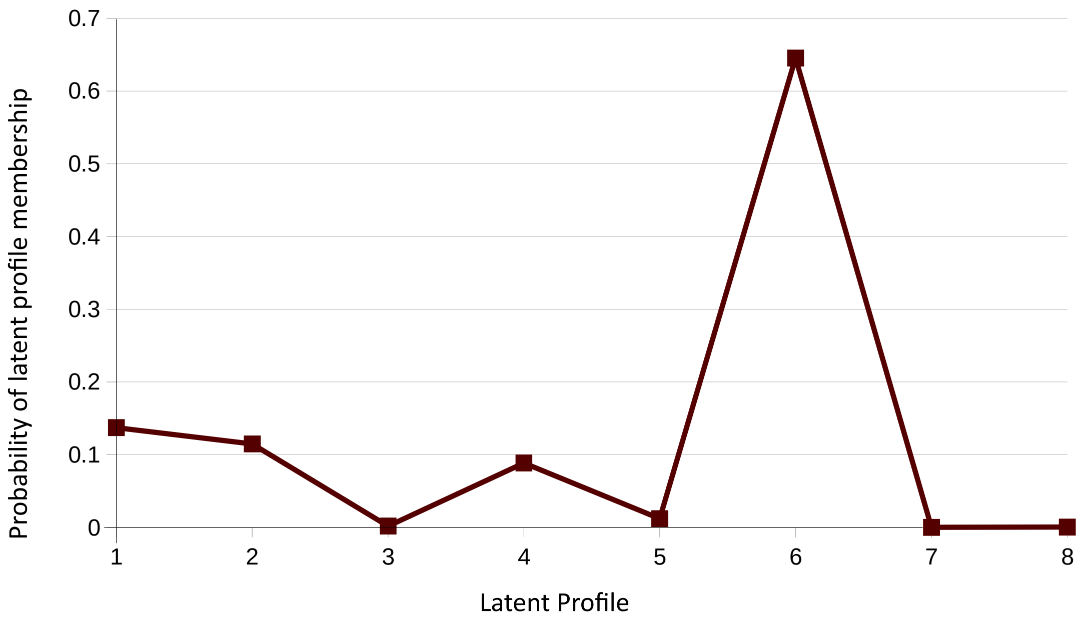


Figure 8: The posterior probability distribution for latent profile membership of Participant #27946

Participant #27946 has the highest chance of belonging to LP6, *p*(LP6)=0.65. The next most probable LP is 1, *p*(LP1)=0.14. They also have a small chance of belonging to any other profile, although this is close to zero for some. MAP translates the probability of belonging to LP6 as a certainty, *p*(LP6)=1.0 ∴ LP_{MAP}=6, disregarding the other PPs and thereby introducing some bias. ML uses all the PPs to include the standard error for profile membership in its calculations, and concludes LP6 is the most likely having taken these into account, *p*(LP6)=0.65, LP_{ML}=6. The mBCH approach is similar to ML, and uses the Participant's data to weight the calculations, which suggests the LP is 4, *p*(LP6)=$0.65_{(weightA)}$, *p*(LP4)=$0.09_{(weightB)}$, LP_{mBCH}=4.

MAP, mBCH, and ML techniques all use statistical likelihoods to calculate the single most appropriate LP. However, potentially more information could be extracted from profile membership by examining the structure of PPs to identify 'secondary profiles' or subgroupings. These can be thought of as the next most probable LP after the primary allocation based on posterior probabilities. In the example above, the Participant has the highest chance of belonging to LP6, and the next highest of belonging to LP1. We can call this profile LP6.1. It is possible other people classified as LP6.1 would have more in common than those classed as LP6.4, even though they all have the same primary profile of LP6. The predictive accuracy of an algorithmic model may therefore be improved by examining patterns within secondary profiles (SP).

It is currently unknown whether there is potential value in refining MAP with +PP classifications (mBCH, ML, or SP) to improve accuracy when associating LPs with treatment outcomes.

8.2. Allocation as usual

IAPT services are not presently mandated to use specific treatment algorithms, and allocation is based most commonly on clinician judgement. Saunders et al. (2016) suggest using the existing MAP algorithm could improve treatment outcomes by increasing appropriate assignment of therapy to different LPs. However, in the IAPT services tested by Saunders and the present study the recovery rate has been improving naturally for a number of years, and was up by 14% in the three years to 2018 (reference omitted to preserve the anonymity of the research site). Improved training and assessment procedures could have potentially contributed to this through better 'intuitive' allocation of appropriate treatment, i.e. allocation patterns closer to that recommend by the algorithm. It is unknown whether allocation as usual (AAU) today is superior to algorithm-derived assignment, given the improvements in the Services' recovery rates since the publication of Saunders et al. in 2016.

8.3. The current study

The study sought to determine whether:

1) the existing profiling method can be improved by incorporating posterior probability information into treatment allocation decisions, and
2) whether this prediction is likely to improve outcomes when compared to standard clinician allocation of treatment.

This occurred in three stages. The first stage used Saunders' et al. (2016) original profiling tool and different methods of incorporating posterior probabilities to see if this improved the accuracy of predictions. The second stage investigated if historical improvements in IAPT recovery rates can be accounted for by the natural allocation of more appropriate therapies to LPs in addition to potential service-related changes such as reduced waiting-lists. Finally, treatment allocation performances using Saunders' original model, a model of enhanced posterior probability inclusion, and clinician allocation as usual were compared to test which is likely to be most effective at influencing treatment outcomes if used today.

The current study will answer:

1) Does inclusion of posterior probabilities improve algorithm accuracy as defined by:
    a) a stronger relationship between latent profiles, intensity of treatment, and treatment outcomes?
    b) prediction of therapeutic outcomes?
    c) prediction of drop-out rate?
2) Based on answers to question (1), do any changes need to occur to the original algorithm's recommendations, particularly regarding which profiles or sub-profiles are likely to benefit from IAPT treatment?

3) Can historical improvements in IAPT recovery rates be accounted for by more appropriate allocation of treatment to the different profiles, and other service-related variables?
4) Which allocation method is likely to be the most effective in IAPT today, between allocation as usual, the original algorithm, and with the inclusion of posterior probabilities?

9. Method

9.1. Research design

This is a secondary analysis of existing IAPT data sets from two joined services. It uses a quasi-experimental design to retrospectively assign participants to MAP or +PP conditions and compare to AAU in order to explore the impact of different allocation models on service-user outcomes. Potential naturalistic changes in LP allocation are explored using multiple time points in a correlational design. Different +PP models are explored in a simulation study design.

9.2. Participants and setting

Participants were clients receiving treatment at two London IAPT services, referred to here as the Services, from 2009-2013 and in 2016. Further participant specification was avoided to increase the generalisability of findings to a typical IAPT setting. All participants previously agreed for their data to be used in research.

9.3. Measures

The IAPT MDS was used to calculate profile membership, which includes nine items relating to demographic details and psychological symptoms (see Figure 9 below and Appendix H).

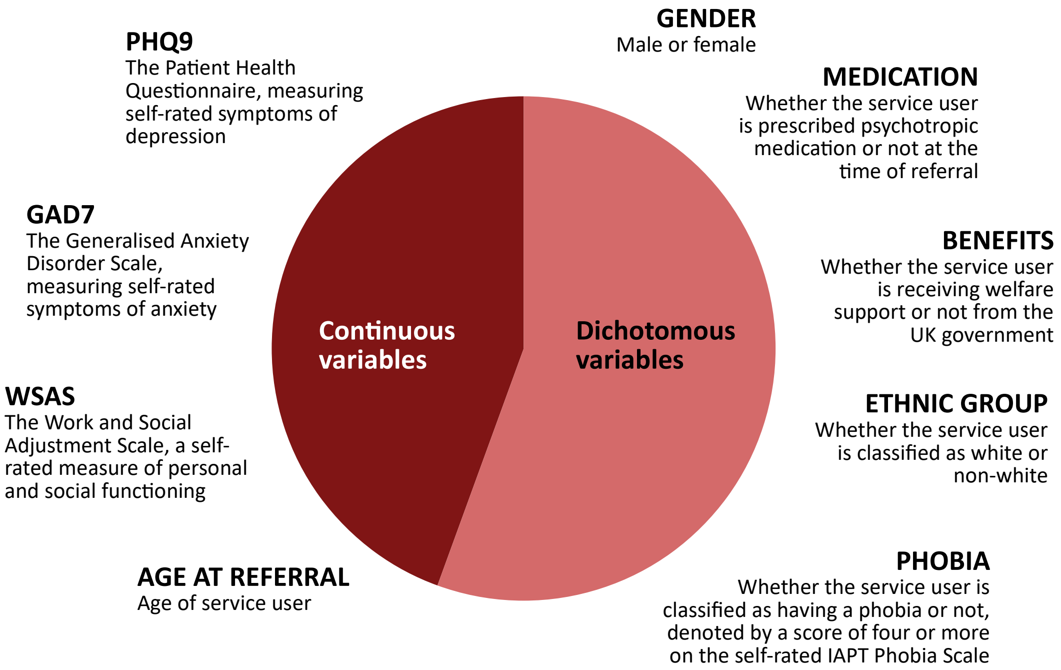

PHQ9
The Patient Health Questionnaire, measuring self-rated symptoms of depression
GAD7
The Generalised Anxiety Disorder Scale, measuring self-rated symptoms of anxiety
WSAS
The Work and Social Adjustment Scale, a self-rated measure of personal and social functioning
AGE AT REFERRAL
Age of service user
Continuous variables
Dichotomous variables
GENDER
Male or female
MEDICATION
Whether the service user is prescribed psychotropic medication or not at the time of referral
BENEFITS
Whether the service user is receiving welfare support or not from the UK government
ETHNIC GROUP
Whether the service user is classified as white or non-white
PHOBIA
Whether the service user is classified as having a phobia or not, denoted by a score of four or more on the self-rated IAPT Phobia Scale

Figure 9: Measures on the IAPT Minimum Data Set used in the current study as observed variables

MDS measures were available for service users' first and last contact with the Services. Changes in PHQ9 and GAD7 scores between first and last contact were used to calculate therapeutic outcomes including reliable recovery, reliable improvement and reliable deterioration, following their respective definitions in IAPT (2014; see Appendix I). Additional data were available on reason for end of treatment (e.g. dropped out, completed a course of treatment, etc.), number of treatment sessions, and intensity of therapy. Service-level variables related to recovery and improvement following Clark et al. (2018) included: proportion of cases with a problem descriptor, mean number of treatment sessions attended, proportion of referrals entering and receiving a course of treatment, mean time waited to start treatment from receipt of referral, proportion of appointments missed, and Index of Multiple Deprivation at the Clinical Commissioning Group (CCG) level.

9.4. Analysis

The analysis is split into three stages. Stage One (Figure 10) established the best algorithmic model to compare against AAU; Stage Two (Figure 11) examined service-level variables linked to AAU; and Stage Three (Figure 12) compared models' performance on therapeutic and service outcomes.

9.4.1. Stage One: comparison of posterior probabilities models

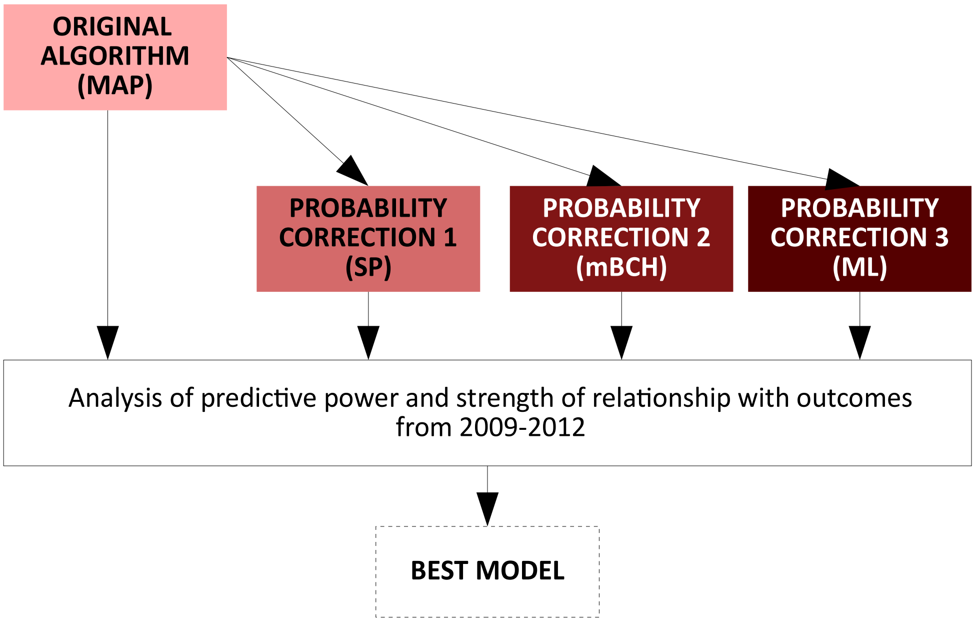

ORIGINAL ALGORITHM (MAP)
PROBABILITY CORRECTION 1 (SP)
PROBABILITY CORRECTION 2 (mBCH)
PROBABILITY CORRECTION 3 (ML)
Analysis of predictive power and strength of relationship with outcomes from 2009-2012
BEST MODEL

Figure 10: Overview of Stage One analysis leading to the selection of a single algorithmic model

Three models correcting for uncertainty of profile membership were compared to the existing MAP model: a modified Bolck, Croon, and Hagenaars (2004) analysis (mBCH; Vermunt, 2010), a Maximum Likelihood approach (ML; Vermunt, 2010), and a secondary profile exploratory investigation (SP). All were derived from the MAP algorithm. ML and mBCH corrections were run with the software package Latent GOLD (Vermunt and Magidson, 2016). SP was explored by first splitting participants into three bands of probability of primary LP membership: high, medium and low. The high probability band had a probability of primary LP membership equal to or greater than 0.9, where $P(LP_n|Y) \geq 0.9$ [the likelihood of achieving that particular LP, n, given that individual's data, Y, is equal to or greater than 90 out of 100 cases]; the medium band where $0.7 \leq P(LP_n|Y) < 0.9$; and the low band where $P(LP_n|Y) < 0.7$ (based on Nagin, 2005). Those in the high probability band were treated as belonging to a single LP with no subprofile, e.g. LP4.0, LP5.0. The medium and low probability bands were classified as having a primary and secondary profile, where the SP is the next most probable LP, e.g. LP5.6, LP2.7. A series of regression analyses were run to determine if treatment outcomes were significantly different between subprofiles with the same primary LP, e.g. between LP1.0, LP1.2, LP1.3...LP1.8. Non-significantly different subprofiles were dropped as categories for subsequent analyses. If the SP model performed well in Stage One, differences between high, medium, and low probability bands would be investigated further.

The models were applied to Saunders' original 2009-2013 dataset for greater comparative validity. The strength of the relationship between LPs and treatment outcomes [research question 1a.] was assessed using an odds ratio effect size calculation and logistic regression to differentiate between high- and low-intensity therapy. Prediction of treatment outcomes by the four models for each LP was compared for accuracy against the original dataset using Bayesian Information Criterion (BIC) scores, pseudo R^2, and percentage of correct predictions [questions 1b.-c.]. Treatment recommendations based on results were devised [question 2].

9.4.2. Stage Two: explaining service variation in rates of reliable recovery and improvement

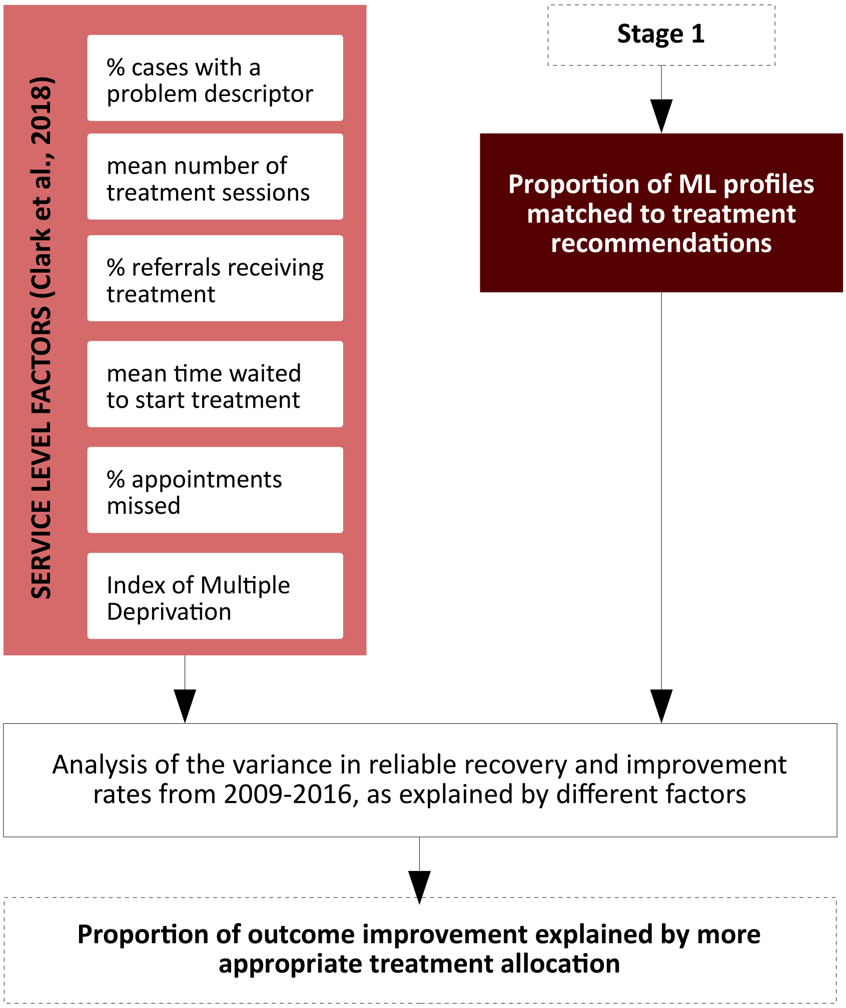

SERVICE LEVEL FACTORS (Clark et al., 2018)
% cases with a problem descriptor
mean number of treatment sessions
% referrals receiving treatment
mean time waited to start treatment
% appointments missed
Index of Multiple Deprivation
Stage 1
Proportion of ML profiles matched to treatment recommendations
Analysis of the variance in reliable recovery and improvement rates from 2009-2016, as explained by different factors
Proportion of outcome improvement explained by more appropriate treatment allocation

Figure 11: Overview of analysis for Stage Two leading to greater understanding of service-level change in recovery and improvement

Historical improvements in IAPT recovery rates [question 3] were examined using Saunders' original 2009-2013 data set and a newer set from 2016 at the same IAPT site. Annual LP distribution within the sets was defined using the most successful algorithm from Stage One, and annual recovery rates for each LP determined descriptively. Other service-related factors were controlled for following Clark et al. (2018) using publicly available data. These factors were: percent of IAPT cases with a problem descriptor, mean number of treatment sessions, percent of referrals receiving treatment, mean time waited to start treatment, percent of appointments missed, and local Index of Multiple Deprivation. These were combined into a single variable for each year by multiplying the factor value for that year by its beta coefficient listed in Clark et al. (2018) and summing the resulting six products. The variation explained by the proportion of LPs receiving algorithm-recommended treatment and Clark et al.'s combined service factors was investigated using linear regression.

9.4.3. Stage Three: finding the most effective method of allocation

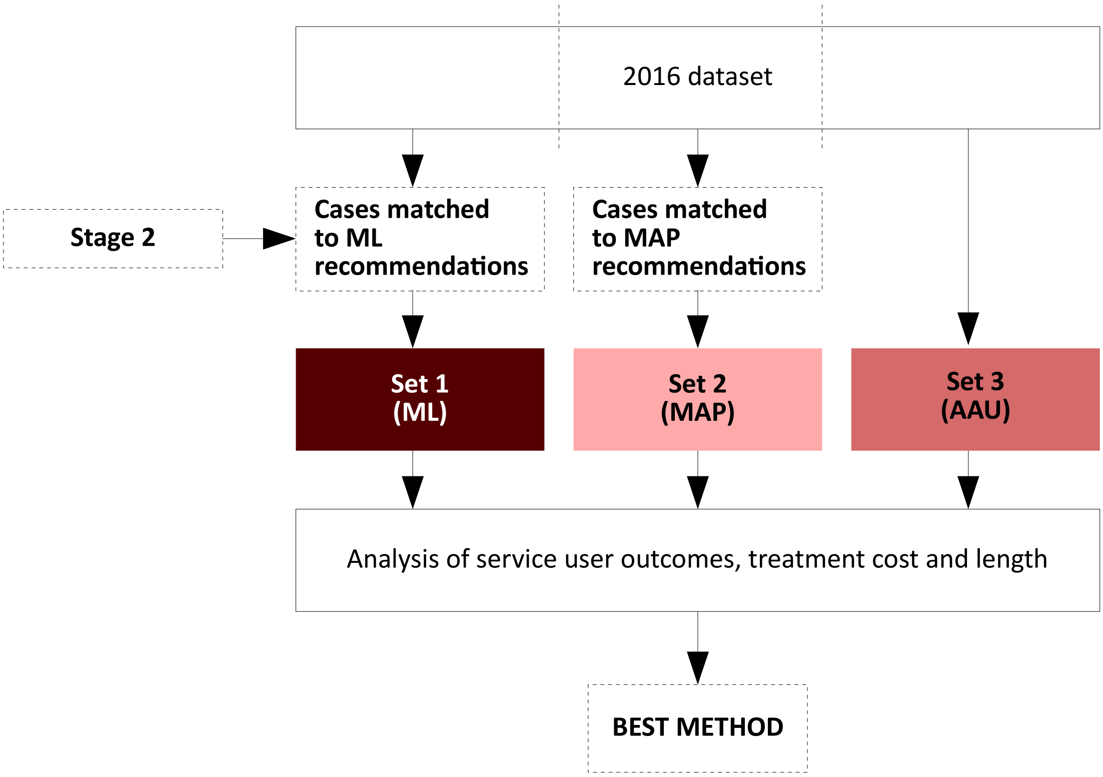

2016 dataset
Stage 2
Cases matched to ML recommendations
Cases matched to MAP recommendations
Set 1 (ML)
Set 2 (MAP)
Set 3 (AAU)
Analysis of service user outcomes, treatment cost and length
BEST METHOD

Figure 12: Summary of Stage Three analysis identifying the most effective method of allocation

Finally, a retrospective comparison of allocation methods on treatment outcomes [question 4] was made using the 2016 dataset between the two most effective algorithms and allocation as usual. The dataset was randomly sampled to form three sets. The first two sets were filtered to include only cases 'matched' to recommended treatment. Those who received the same treatment their respective algorithm would have recommended were considered 'matched'; cases receiving a higher intensity than recommended were 'over-matched'; and receiving a lower intensity was 'under-matched'. The AAU set was not filtered. The sets were then homogenised using case-control matching, and a one-way ANOVA used to compare therapeutic outcomes, treatment length, and treatment cost.

9.5. Ethical considerations

All data was anonymised at or before the point of collection by the original collecting researcher. All study data was subject to data handling safeguards and stored in accordance with the Data Protection Act (Her Majesty's Stationery Office, 1998). All participants previously agreed for their data to be used in research. This research was granted ethical approval by the Research Department of Clinical, Educational and Health Psychology at University College London.

10. Results

10.1. Missing data bias

A missing data analysis was performed on the pooled MDS assessment data. Three items had more than 5% of their data missing: benefits status (40.2% missing), WSAS (12.5%), and medication status (11.0%). An EM imputation with Little's MCAR test (Chen & Little, 1999) indicated data were not missing completely at random, $\chi^2(21, 19917)=285.25$, $p<0.05$. A pattern analysis suggested WSAS and benefit status were missing together in 16% of cases, and medication and benefit status were not present together in 10% of cases. It was not possible to determine whether data were missing not at random. This is further discussed in the Strengths and Limitations section on page 112.

10.2. Stage One: Incorporating posterior probabilities into the original algorithm

Of the 33,363 cases in the 2009-2013 dataset, 18,023 were removed due to incomplete outcome data or not entering IAPT at caseness (n=15340). To create the SP model, cases were divided based on their posterior probabilities as described in the Methods section on page 94. For example, if a case has the highest probability of belonging to MAP profile 1 and the next highest as MAP profile 2, it is referred to as SP1.2. The groupings for SP3.6 and 6.3 had too few cases (<10) for analysis and were removed by reallocating them to their respective highest-probability profiles. A one-way ANOVA with Bonferroni correction was then used to compare all the SPs for significantly different outcomes on reliable recovery; those not significantly different to other SPs in their class (e.g. between SP4.1, SP4.2,...SP4.8) reverted to their original LP. These analyses were repeated until only statistically distinct SPs remained, of which there were four in additional to the other eight profiles: 1.2, 1.6, 2.6, and 3.4.

10.2.1. Research Question 1: Does inclusion of posterior probabilities improve algorithm accuracy?

Logistic regression analyses of treatment outcomes (reliable recovery, reliable deterioration, etc.) predicted by model and therapy intensity, and split by profile suggested the ML model had the strongest overall relationship between treatment outcomes [research question 1a.]. MAP and ML profiles differed in 7% (n=2337) of cases. A summary of MAP and ML profile characteristics and outcomes are given in Appendix J on page 205.

In a set of logistic regression analyses where model was the only predictor, profile significantly predicted treatment outcomes regardless of which model was used, $p<0.05$ (a regression was run three times for each model, corresponding to the three treatment outcomes as dependent variables). However, when therapeutic intensity was introduced the models differed in the number of specific profiles associated with each outcome. This was found in a separate set of logistic regression analyses where model and intensity were predictors and the cases were split by profile. A regression was run for each treatment outcome and model (i.e. the three outcomes x the four models, or 12 runs). As the cases were further split by profile, each run included a regression for each profile, i.e. eight or 12 regressions per run. This indicated which profiles were affected by therapy intensity, and the strength of this effect could then be compared between models. When choice of treatment intensity had significantly greater odds of improving outcomes for a particular profile, the number of people potentially impacted was calculated and used in determining the strength of the relationship. For comparison, the results for the MAP and ML recovery analyses are summarised below in Table 6, and the full output can be seen in Appendix K.

Table 6

Comparison of MAP and ML profiles with significant odds ratios (OR) of achieving reliable recovery (RR) at different intensities of therapy, from clients starting treatment at caseness

	MAP				**ML**			
	Number of cases (% RR)				**Number of cases (% RR)**			
Profile	**Low-intensity**	**High-intensity**	**OR***	**Additional RR cases****	**Low-intensity**	**High-intensity**	**OR***	**Additional RR cases****
1	-	-	-	-	733 (44%)	341 (50%)	1.31	44
2	2440 (45%)	1038 (51%)	1.31	149	2191 (45%)	919 (51%)	1.29	131
6	793 (32%)	608 (42%)	1.54	79	867 (33%)	638 (43%)	1.50	87
7	797 (17%)	1195 (20%)	1.27	24	-	-	-	-
8	2048 (31%)	1180 (35%)	1.23	82	1969 (31%)	1104 (36%)	1.25	98
Sample total (% RR)	8003 (36%)	5196 (37%)		332	8003 (36%)	5196 (37%)		361

*Significant at p<0.05. Where OR>1, recovery odds are greater with high-intensity. Where OR<1, recovery is more likely with low-intensity

**If all profile members in the preceding row received the most effective intensity of therapy as indicated by their OR

An odds ratio greater than one suggests profile members are significantly more likely to recover if they receive high-intensity therapy compared to low-intensity. In MAP, these were profiles 2, 6, 7, and 8. In ML, which calculates its profiles slightly differently, profiles 1, 2, 6, and 8 had improved odds. If sample members of the relevant profiles all received their most effective intensity, we could expect 332 additional cases of recovery in MAP compared to AAU, or 361 additional cases in ML compared to AAU. Similar analyses were performed on the other outcomes.

The results from these analyses for reliable recovery, reliable change, and deterioration are summarised in Table 7 below. Each outcome was most closely associated with a different model. The original algorithm was outperformed on every outcome by an alternative model, although still performed well overall.

Table 7
Summary of model performances when sample profiles receive their recommended therapy intensity

	Additional reliable recovery cases	**Additional reliable change cases**	**Additional cases not deteriorating**	**Total additional cases benefiting***	**Benefit as percent of sample**
MAP	332	63	8	403	3.06%
SP	256	105	22	383	2.88%
mBCH	189	158	18	365	2.76%
ML	361	61	19	410	3.10%

*Maximum possible given a forced choice of intensity

In one case (ML profile 8) the indicated therapeutic intensities for maximising recovery and minimising deterioration conflicted, meaning the intensity chosen would result in either less recoveries or more deteriorations than usual. The Total Additional Cases Benefiting column thus reflects the maximum possible number of people who could gain from the model (and not the sum of the row), in this case taking into account a higher than AAU deterioration rate. Despite this, ML profiles had the strongest relationship to intensity and outcome overall as gauged by number of people benefiting.

Model fit [research question 1b.] was estimated using BIC scores, pseduo-R^2, and percentage of correct predictions, using the previous logistic regressions where model was the only predictor and cases were not split (a total of 12 regressions). BIC was chosen as it is more reliable with large data samples (Kieseppä, 2003), and Nagelkerke's pseudo-R^2 was chosen as it more closely resembles typical R statistics, including being standardised for easier comparison between models (Nagelkerke, 1991). Results are given in Table 8.

Table 8
Performance of models as predictors of treatment outcomes, based on clients starting treatment at caseness

Outcome	**Model**	**BIC**	**Significance**	**Percentage of correct predictions**	**Nagelkerke's pseudo-R2**
Reliable recovery	MAP	134.84	<0.001	63.5%	0.055
	SP	208.88	<0.001	63.5%	0.062
	mBCH	135.38	<0.001	63.4%	0.055
	ML	135.62	<0.001	63.4%	0.055
Reliable change	MAP	135.30	<0.001	60.3%	0.020
	SP	209.50	<0.001	60.3%	0.029
	mBCH	135.87	<0.001	60.3%	0.025
	ML	136.05	<0.001	60.2%	0.020
Reliable deterioration	MAP	125.79	<0.001	92.5%	0.041
	SP	194.65	<0.001	92.5%	0.044
	mBCH	126.47	<0.001	92.5%	0.044
	ML	126.30	<0.001	92.5%	0.040

BIC values closer to zero or a pseudo-R^2 closer to one suggest a better fit. The low pseduo-R^2 values here, unlike Pearson's R^2, are more typical in logistic regression and are not in themselves suggestive of poor fit (Hosmer, Lemeshow, & Sturdivant, 2013); the statistic is more useful for directly comparing models. In terms of goodness of fit, MAP more accurately matches the data based on BIC and percent of correct predictions. SP is less accurate, although has the highest pseduo-R values indicating a better relative fit (compared to an intercept model). Combined with the previous analyses, this leads to the intriguing conclusion that MAP is the best prognostic model, whereas ML is the best model for informing decisions on selecting treatment.

When calculating dropout rates [research question 1c.] it was important to first establish whether this would be useful information to judge the models on. A frequency count of sample members entering at caseness and with T2 data showed that 59.3% of clients reliably recovered when they completed a course of therapy. No one recovered before they dropped out, suggesting reducing dropout rates is an important way to maximise potential benefit from therapy.

As before, a logistic regression was performed for each model to determine whether intensity of intervention improved the odds of dropping out of treatment. Statistically significant profiles were examined to calculate the total number of people who could be affected by specifying therapeutic intensity. These were then weighted by their respective profiles' deterioration, and recovery and reliable change rates, as summarised in Table 9 below.

Table 9

Impact of models on dropout rate when most effective therapeutic intensity is used, based on clients starting treatment at caseness

Model	Additional cases completing treatment	Estimated additional cases achieving reliable recovery or change	Estimated additional cases deteriorating	Estimated additional number of cases benefiting
MAP	472	297	40	257
SP	487	304	42	262
mBCH	492	316	39	277
ML	453	285	39	246

All figures are to the nearest whole number

Using the mBCH potentially benefits the most clients by reducing dropout rates. However, these results should be interpreted with caution as it is not known whether people who drop out have moderating characteristics affecting treatment efficacy compared to other groups. The rates used to calculate recovery and deterioration were based on the profile as a whole, and may be higher or lower for those who drop out. For this reason, these estimates were not totalled with the previous benefit calculations in Table 7 but were instead considered alongside them.

These analyses suggest the ML and MAP models are the most usefully accurate algorithms. Although MAP had superior predictive power, its relationship to outcomes and therapeutic intensity was not a strong as ML. Factoring in dropouts gives MAP an advantage, potentially impacting 5% of the sample (660 people) compared to ML at 4.97% (656). However, the unknown reliability of the dropout benefit estimates gives greater weight to the recovery, reliable change and deterioration calculations, arguing ML could be the most appropriate choice of model. As the results are so comparable, both algorithms will be tested alongside Allocation As Usual in Stage Three. Overall, incorporating posterior probabilities improved algorithm accuracy for several outcomes, but choice of outcome priority is vital to determine effectiveness as no one model was consistently better across all of them.

10.2.2. Research Question 2: Does the original algorithm need updating?

ML profiles had greater odds of recovery at different intensities compared to MAP, so the original recommendations need modifying. For transparency, the assumptions upon which the recommendations are based are included in Appendix L. Recommendations were informed by an additional analysis on the impact of ‘stepping’ ML profiles ‘up’ to a higher intensity therapy or ‘down’ to a lower intensity following a dose of therapy, which is included in Appendix M. The recommendations in Table 10 below are mainly based on whether the odds of achieving reliable recovery are significantly increased with a particular therapeutic intensity.

Table 10

MAP and ML treatment recommendations for profiles

Profile	Original MAP Recommendation	ML Recommendation
1	Initiate at Step 2, high probability of recovery	Initiate at Step 3, moderate* probability of recovery. Likely to benefit from step up if starts at Step 2
2	Initiate at Step 2, monitor until session 3	Initiate at Step 3, moderate probability of recovery
3	Initiate at Step 2, high probability of recovery	Initiate at Step 2, moderate probability of recovery
4	Initiate at Step 2, moderate probability of recovery and unlikely to benefit from step up	Initiate at Step 2, lower probability of recovery
5	Initiate at Step 2, lower probability of recovery and unlikely to benefit from step up	Initiate at Step 2, lower probability of recovery
6	Initiate at Step 3, moderate probability of recovery	Initiate at Step 3, lower probability of recovery. Likely to benefit from a step up if starts at Step 2
7	Unlikely to benefit from IAPT service, specialist service recommended	Initiate at Step 2, unlikely to recover, moderate chance of reliable change
8	Initiate at Step 2, lower probability of recovery and step up may increase probability of recovery	Initiate at Step 3, lower probability of recovery

*Moderate: 50%≤p<75%; Low: 25%≤p<50%; Unlikely: p<25%

These recommendations will be used for the analysis in Stage 2, and to assign profiles to their designated intensities for the comparison with AAU in Stage Three.

10.3. Stage Two: Historical improvements in IAPT recovery rates

In Clark et al. (2018), reliable recovery and reliable improvement rates for services could be predicted by six annual variables, including mean number of treatment sessions, percent of appointments missed, and local Index of Multiple Deprivation (IMD; Noble, Wright, Smith, & Dibben, 2006). At the time of Clark’s study these data were freely available on NHS Digital via the IAPT Data Set, which has since closed. Data points for this study have thus been gathered from multiple sources, including IAPT Annual Reports, NHS data files, gov.uk, and the study data sets. This may cause some variability between figures reported here and what could be obtained elsewhere, but in every case the best available evidence has been used.

To test whether recovery and improvement at the service level could be linked to the variability in the proportion of profiles receiving their most effective treatment intensity each year, the two study datasets were combined. Annual averages for the relevant factors (recovery, number of treatment sessions, etc.) were calculated or entered from other sources for the six years of 2009-2013 and 2016. ML profiles were classified according to the algorithm in Stage One, and treatment received was compared to the recommendations in Table 10 for congruence. The proportion of clients receiving the recommended treatment intensity for each year was then calculated. Annual IMD scores—which in Clark were used at the individual CCG level—were averaged between the various CCG areas covered by the Services. Due to the relatively small number of observations at the service level and a higher chance of overfitting, Clark's variables were combined into a single value for each year as described previously. Both factors were modelled using linear regression to compare the effect adding allocation accuracy as a factor to Clark's model. To check for overfitting the analysis was re-run, this time adding Clark to allocation accuracy to see if the same result was achieved.

10.3.1. Research Question 3: What best accounts for historical improvements in recovery?

Reliable improvement rate was neither significantly predicted by Clark's model alone nor with the addition of allocation accuracy. Reliable recovery rate was significantly predicted, the results for which are shown in Table 11 below.

Table 11
Linear regression for reliable recovery comparing predictive variability of Clark et al.'s model with and without therapeutic allocation accuracy

Model	**R**	**Adjusted R^2**	**R^2 change**	**F change**	**F change significance**
Clark et al. 2018	0.729	0.415	0.532	4.549	0.100
Clark et al. 2018 + Allocation Accuracy	0.966	0.899	0.401	18.040	0.024*

*R^2 is significant

The model including both allocation accuracy (Beta = 0.60, Standard error = 0.14, *p*<0.05) and Clark's variables (B = 0.023, SE = 0.008, *p* = 0.072) significantly improved the predictive ability of the model over Clark alone, accounting for 90% of the variance in reliable recovery scores. Recovery and allocation accuracy appeared highly correlated, as shown opposite in Figure 13. Overfitting was not indicated, however due to the small number of observations (n=6) the above results should still be interpreted as exploratory. A visual inspection of Figure 13 suggests there may be a 'true' relationship between allocation accuracy and recovery rates, but the precise nature of that relationship can only be suggested at this point.

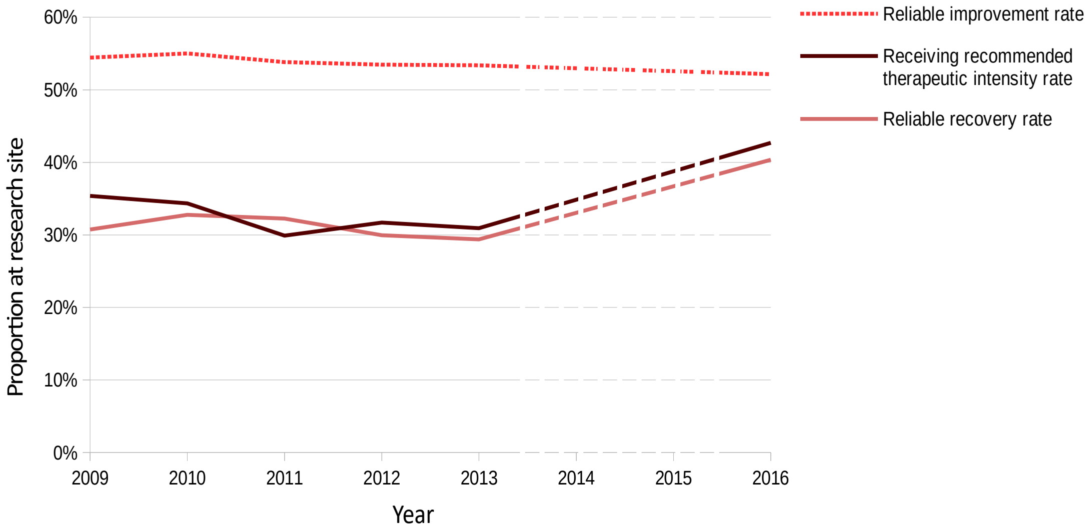


Figure 13: Percentages of clients achieving reliable recovery or improvement, and percentage of profiles allocated to the recommended therapeutic intensity for selected years. Note: data points for 2014-15 are extrapolated

The regression results and above chart suggest historical improvements in reliable recovery can be partly accounted for by a natural improvement in allocation accuracy. Approximately ten percent of the variation is unaccounted for, which could be attributable to factors including more effective therapies or staff morale. Reliable improvement at the Services does not seem to be linked to service factors or allocation accuracy—or, curiously, recovery rate. Improvement rates appear roughly stagnant over the measured period.

10.4. Stage Three: Comparison of allocation methods

The 2016 dataset was divided into three groups using a random number generator (The Document Foundation, 2018) to assign each case a number between one and three. These corresponded to the ML, MAP, and AAU groups respectively. The ML and MAP groups were further selected for those profiles matched to their respective treatment recommendations (based on Table 10 from page 106). A one-way ANOVA with Bonferroni Correction was performed to check homogeneity of MDS variables between all groups, of which five were significantly different. Of these, all but phobia, GAD7 and WSAS scores were negligibly different. Three-way case-control matching against these variables was achieved using MedCalc (MedCalc, 2018), resulting in a final homogenised sample of 2854 (ML=797, MAP=674, AAU=1383).

10.4.1. Research Question 4: Which allocation method is most effective?

A one-way ANOVA with Bonferroni Correction comparing the impact of allocation method against reliable recovery, reliable change, and reliable deterioration as binary outcomes was conducted. There was a significant effect of allocation on recovery between ML (mean[SD]=0.37[0.48] and MAP (0.45[0.5]), F(2,2851) = 4.21, *p*<0.05. All other interactions were non-significant. This suggests MAP is a more effective allocation method for treatment outcomes than ML, but is no different to AAU.

Exploring this result further, a similar analysis found no effect of allocation method on number of treatment sessions, F(2,2851)=0.93, *p*>0.05, but did for treatment cost, F(2,2851)=34.66, *p*<0.05. When average cost of treatment was compared MAP was significantly cheaper than ML or AAU, where the mean course of treatment was £616.32 (SD=444.8), £976.90 (SD=879.0), and £799.45 (SD=714.9) respectively. Cost was calculated according to the per session figures provided by Radhakrishnan et al. (2013). These specific values are for reference only and are not meant to stand in for more thorough and valid calculations using appropriate health economic methodologies, which regrettably could not be undertaken here due to time constraints.

The difference in cost but not number of treatment sessions can be understood in terms of the proportion of clients allocated to high-intensity therapy, which was significantly different between models, F(2,2460) = 142.49, *p*<0.05. ML had the highest proportion of services users allocated to high-intensity (61%), followed by AAU (46%), then MAP (7%). This is similar to the proportion of clients 'over-matched' (allocated a higher intensity than suggested) when considering MAP recommendations: ML over-matched 61% of profiles, and AAU over-matched 47%.

10.5. Summary: Which model is most likely to improve outcomes today?

Comparing the three different allocation models suggests there is today little difference between them when considering outcomes for service users—namely reliable recovery, improvement, or deterioration. Analysis of historical trends suggests this is because IAPT clinicians are naturally allocating more appropriate therapy to profiles. However, when the cost of treatment is considered, using the MAP algorithm is significantly cheaper as it is less likely to recommend an unnecessarily high intensity.

11. Discussion

Using posterior probabilities in different models of profile allocation did not enhance therapy outcomes compared to the original MAP algorithm. Utilising MAP to recommend treatment intensity instead of using allocation as usual could benefit clients by assigning a shorter and equally effective therapy. This fits with the existing literature where relatively simple models are able to out-perform more experienced clinicians. However, the algorithm cannot be seen as an effective way to improve recovery or reliable change rates at this particular IAPT site today, as clinicians are naturally matching more profiles to their most appropriate intensities than before. How this change occurred, and in a relatively short period of time, are interesting questions. One explanation could be a growth in the collective experience at the Services: as more clients are seen, clinicians become more adept at intuitively knowing what works for which groups and become better at passing on that knowledge—a phenomenon also known as ‘chicken sexing’ (Horsey, 2002). IAPT is a relatively new initiative, graduating from a pilot site in 2008, meaning many services have been running only a handful of years. It would be useful to see if similar patterns are observed elsewhere, and whether there is a ceiling effect where the collective knowledge reaches a natural saturation point after so many years of operation.

One intriguing finding was reliable improvement rates were not affected as much as recovery rates, and at the Services appeared stagnant over the measured period despite increases in recovery. This could be due to a slight increase in average clinical scores between 2009-2013 and 2016 of approximately 1.5 points on both the PHQ9 and GAD7 scales. With more people entering the Services at caseness, more clients will register on the recovery statistics even though the improvement rates remain the same. This also offers another explanation for the improvement in allocation rates: clinicians could just be better judges of therapeutic intensity for clinical cases over non-clinical. Anxiety and depression scores were not investigated in Clarke’s analysis of service-level variables, and are worth exploring further in future models.

11.1. Strengths and limitations

To the best of the author’s knowledge this research is the first to simulate different ways of modelling algorithmic processes for the purpose of improving therapeutic allocation. Its conclusions are strengthened by comparing recommendations at different time points, which demonstrated a change in service-level factors that impacted the evaluation of its usefulness. Therefore a key strength of this research is also a significant limitation, and justifiably so. The results apply to a specific service in a particular time frame, and will not have the same utility outside of this. The findings are particularly sensitive to service variables, which as we have seen can change considerably in a few years. In a crude comparison of the average cost of treatment, using MAP could have saved around £330 per treatment in 2013 and improved recovery rates, but only £180 in 2016 with no impact on recovery[3]. Applying the algorithm to other IAPTs should be similarly variable. However, as long as a sufficient amount of relevant data is used to update the algorithm with appropriate frequency, the same research process should produce a valuable tool regardless of setting.

The importance of updating the algorithm is demonstrated in this research by the recommendation attached to Profile 7. In Saunders' original analysis only 22% of Profile 7s improved, leading to the recommendation not to treat this profile at IAPT on the basis they were unlikely to benefit. Profile 7s were therefore largely excluded from the Stage Three MAP group[4] as the only way to match them to treatment was to refer to another service. However, in Stage One 50% of MAP Profile 7s achieved reliable improvement, arguing they do benefit from treatment even though the chances of recovery remain low. It then follows that MAP recommendations should be updated to suggest Profile 7 is initiated at low-intensity therapy (and then potentially step up) with a moderate chance of benefiting. Treating Profile 7s would reduce the overall recovery rate and increase costs as reported here, although still result in a significant overall saving compared to AAU. It is also worth bearing in mind much of the analysis and the profiles themselves were based on clients entering IAPT at caseness, and so would not have the same applicability to service users without clinically significant depression or anxiety. However, this does not affect reliable recovery outcomes as these already exclude clients not at caseness.

[3] Average calculated cost of treatment based on AAU from 2009-2013 was £949.44

[4] A separate analysis was run excluding Profile 7 from the ML model as well to see if this had confounded the results, but it did not change the conclusion

It would have been useful to attempt a re-validation of the various model MAP profiles using the newer dataset through out-of-sample forecasting. This would have provided further evidence for or against the reliability of the profile groupings, although time restraints limited the opportunities for testing. Saunders reports using a validation sample while developing MAP in his original paper, suggesting the MAP profiles are relatively stable.

The reliability of the findings may be affected by a potential missing data bias. The analysis in Section 10.1 suggested data were not missing at random, arguing the profiles may not represent the full range of IAPT service users. It is encouraging that an improvement in outcomes, especially historic recovery rates, can be seen when allocation is related to the existing profiles. This suggests the profiles are reliable enough, even if data are systematically omitted. However, a large proportion of information on benefit status in particular was missing, and it would be useful to explore this further. The analysis could have been strengthened by a missing data imputation, but this was omitted to improve comparison with the original analysis.

One final limitation involves discussions around the use of race, gender and benefit status in the algorithm. It is beyond the scope of this research to fully examine the philosophical implications of assigning therapy based on biological and social background—especially in one paragraph—but it is nevertheless important to start. The categorisation of ethnicity as ‘White or not-White’ is an interesting example; it suggests either there is something inherent in Caucasians that causes a different reaction at particular therapeutic intensities compared to non-Caucasians, or that it is not measuring race. The algorithm could be detecting prejudice experienced by clients with non-native skin tones as a social determinant of mental health, for example. This a potential limitation to using the MDS variables as done in this study, and warrants further consideration.

11.2. Professional implications

This study argues the algorithm is a valuable tool in IAPT settings and would likely benefit professional practice. Here, the core themes are decisions, context, and data. First, it is important to remember that even though it is a procedural mathematical construct, the final algorithm was produced from a series of subjective, very human decisions that influence its utility. For instance, recovery was prioritised over improvement; change was defined in terms of depression and anxiety not functioning; shorter therapy was favoured over longer-term. Any professional seeking to use this or similar tools would do well to consider what is important to achieve in their setting, what assumptions their decisions are based on, and change their algorithm accordingly.

Second, the recommendations rely on a certain consistency of treatment that is likely unique to the service context. Both low- and high-intensity IAPT therapies normally refer to therapist-led Cognitive Behavioural Therapy, but may also include computerised CBT, mindfulness, Interpersonal Therapy, Behavioural Couples Therapy and Dynamic Interpersonal Therapy. The ratios of particular therapies will differ between services and times, and delivery affected by individual therapists and policies; even manualised CBT is unlikely to be delivered exactly the same way between two services. These factors influence the efficacy of the therapeutic intensity and thus the recommendations for a given setting.

Third, this type of tool is context-dependent but its awareness of that context depends on the data it is fed. Poor quality, patchy, out-dated, or displaced data will reduce its suitability for a given setting. This should be factored in to cost-benefit calculations when considering implementation. IAPT is well-suited to this requirement due to its routine data collection policy, but individual private practice for example may require more substantive changes. In the Services today, the costs of implementation would be outweighed by the treatment savings, but this may not be the case in all services.

11.3. Future research

This research took eight distinct profiles and allocated them to one of two conditions; there is every chance that additional specialisation of treatment may further improve outcomes. For instance instead of offering all the available high-intensity therapies as one group, the least effective ones could be dropped for particular profiles. For example some profiles might benefit from a restricted allocation choice of (low-intensity) group CBT or mindfulness, while others might have greater odds of recovery when allocated to either (high-intensity) computerised therapy or community-based intervention. A profile might have better chances of improving if stepped up to Interpersonal Therapy than individual CBT. Data on the type of therapy offered was not available to this study, but their analysis would make a logical next step in profile-based allocation.

It is unknown whether these profiles identified by Saunders are unique to the Services or are reproducible elsewhere—or even if a covert ninth profile lurks somewhere in the missing data items. Further latent profile analyses in different parts of the UK and elsewhere would make an interesting comparison and hasten investigation into further treatment specialisation if profiles are reproduced. If the profiles are not apparent elsewhere, greater investigative focus on the process of adapting this algorithm to local profiles would then be useful.

11.4. Conclusion

This research shows mental health outcomes and service costs can be improved by using algorithms to allocate clients to therapy based on latent profiles. The extent of this improvement is data- and context-dependent, and invites further conversations about how we make clinical decisions, whether by machine or by human.

12. References

Bayes, T. (1763). An Essay towards Solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society of London*. http://doi.org/10.1093/biomet/45.3-4.293

Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*. http://doi.org/10.1093/pan/mph001

Clark, D. M., Canvin, L., Green, J., Layard, R., Pilling, S., & Janecka, M. (2018). Transparency about the outcomes of mental health services (IAPT approach): an analysis of public data. *The Lancet*, *391*(10121), 679–686. http://doi.org/10.1016/S0140-6736(17)32133-5

Clarke, D. E., Brown, A. M., & Griffith, P. (2010). The Brøset Violence Checklist: Clinical utility in a secure psychiatric intensive care setting. *Journal of Psychiatric and Mental Health Nursing*, *17*(7), 614–620. http://doi.org/10.1111/j.1365-2850.2010.01558.x

Her Majesty's Stationery Office. (1998). *Data Protection Act 1998*. *Legislation.Gov.Uk*. http://doi.org/http://www.legislation.gov.uk/ukpga/1998/29/pdfs/ukpga_19980029_en.pdf

Horsey, R. (2002). The art of chicken sexing. *UCL Working Papers in Linguistics*. Available from http://cogprints.org/3255/1/chicken.pdf

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression: Third Edition*. *Applied Logistic Regression: Third Edition*. http://doi.org/10.1002/9781118548387

Huijbregts, K. M. L., De Jong, F. J., Van Marwijk, H. W. J., Beekman, A. T. F., Adèr, H. J., Hakkaart-Van Roijen, L., … Van Der Feltz-Cornelis, C. M. (2013). A target-driven collaborative care model for Major Depressive Disorder is effective in primary care in the Netherlands. A randomized clinical trial from the depression initiative. *Journal of Affective Disorders*, *146*(3), 328–337. http://doi.org/10.1016/j.jad.2012.09.015

Kieseppä, I. (2003). AIC and large samples. *Philosophy of Science*, *70*(5), 1265–1276. http://doi.org/10.1086/377406

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A New Depression Diagnostic and Severity Measure. *Psychiatric Annals*. http://doi.org/10.3928/0048-5713-20020901-06

Kroenke, K., Spitzer, R. L., Williams, J. B. W., Monahan, P. O., & Löwe, B. (2007). Anxiety disorders in primary care: Prevalence, impairment, comorbidity, and detection. *Annals of Internal Medicine*. http://doi.org/10.7326/0003-4819-146-5-200703060-00004

MedCalc (Version 18.10.02) [Computer software]. Ostend: MedCalc Software

Mundt, J. C., Marks, I. M., Shear, M. K., & Greist, J. H. (2002). The Work and Social Adjustment Scale: A simple measure of impairment in functioning. *British Journal of Psychiatry*. http://doi.org/10.1192/bjp.180.5.461

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika 78*(3), 691-692.

Nagin, D. S. (2005). *Group-Based Modelling of Development*. London: Harvard University Press.

National Health Service England. (2018). *Service standards*. Retrieved from https://www.england.nhs.uk/mental-health/adults/iapt/service-standards/

Noble, M., Wright, G., Smith, G., & Dibben, C. (2006). Measuring multiple deprivation at the small-area level. *Environment and Planning A*. http://doi.org/10.1068/a37168

Nuffield Trust. (2018). *Improving Access to Psychological Therapies (IAPT) programme*. Retrieved January 2, 2019, from https://www.nuffieldtrust.org.uk/resource/improving-access-to-psychological-therapies-iapt-programme

Public Health England. (2018). *Public Health Profiles: Recovery rate for all anxiety and stress related disorders: % of referrals finishing a course of treatment who are "moving to recovery" (annual)*. Retrieved January 2, 2019, from https://fingertips.phe.org.uk

Radhakrishnan, M., Hammond, G., Jones, P. B., Watson, A., McMillan-Shields, F., & Lafortune, L. (2013). Cost of Improving Access to Psychological Therapies (IAPT) programme: An analysis of cost of session, treatment and recovery in selected Primary Care Trusts in the East of England region. *Behaviour Research and Therapy*. http://doi.org/10.1016/j.brat.2012.10.001

Saunders, R., Cape, J., Fearon, P., & Pilling, S. (2016). Predicting treatment outcome in psychological treatment services by identifying latent profiles of patients. *Journal of Affective Disorders*, *197*(0), 107–115.

Stallvik, M., Gastfriend, D. R., & Nordahl, H. M. (2015). Matching patients with substance use disorder to optimal level of care with the ASAM Criteria software. *Journal of Substance Use*, *20*(6), 389–398. http://doi.org/10.3109/14659891.2014.934305

The Document Foundation. (2018). LibreOffice Calc (Version 6.1.1.2) [Computer software]. Available from https://www.libreoffice.org

The National Institute for Health and Care Excellence. (2009). *The NICE Guideline on the Treatment and Depression the Treatment and Management of Depression*. *NICE guidelines [CG90]*.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*. http://doi.org/10.1093/pan/mpq025

Vermunt, J. K., & Magidson, J. (2016). Latent GOLD (Version 5.1) [Computer software]. Massachusetts: Statistical Innovations Inc.

PART THREE—CRITICAL APPRAISAL

VIRTUOUS SOULS AND AUTOMATONS

Looking back on the process of writing this thesis raises a few challenging perspectives I will try to navigate with some care: first, how we position ourselves as humans or algorithms in research and second, whether we use idiosyncratic or rigid methods to do that research. Finally, I examine the thesis in light of itself, using the discussion insights and results from the literature review to critically appraise the research project.

13. Are you a robot?

I will illustrate my first point with writing styles. There are many ways to communicate technical ideas, although as Montgomery (2003) points out, the majority of guides “boil down to collections of rules, standards, and warnings” (p. 4). This is fertile ground for algorithms, opening up academic writing to automation. SCIgen (Stribling, Krohn and Aguayo, 2005) for instance, which auto-generates whole papers, had 120 known published ‘studies’ on SpringerLink Journals (van Noorden, 2014). The following two paragraphs come from Essaybot (2018) and AI-Writer (2018) respectively, generated using only the keywords ‘algorithm’ and ‘psychology’.

> Algorithms are often used in mental health analysis because their nature implies the psychologist will have trust in the results. Checking symptoms, uniformity of results, and effectiveness of interventions can all be defined and shown through the use of specialised algorithms. Let us look at symptoms verification: people who are struggling with depression, anxiety, or some related condition frequently do not know what circumstances produce the emotional reaction. They only realise they often seem sad or nervous. Although the therapist might have some thoughts of the sorts of elements that might be causing their reaction, the algorithm will provide a

> comprehensive set of steps towards verifying the conditions that are causing the depression or anxiety effects.
>
> Behaviour can reflect the interaction of cognitive and emotional factors, and can be captured more effectively by using a focus approach on the interaction between different decision-making systems. For example with heterogeneity in personality types, decision-making in rule-based models (such as bayenda upgrading) helps to ensure consensus between different personalities, promoting effective social decision-making processes despite the natural heterogeneity.

As an example of a formal writing style, they are not bad. The second paragraph is partly gibberish (there is no such thing as bayenda upgrading), yet otherwise they could pass for human. The question is, given algorithms are good enough to write a paper on their own, what do we 'biologics' bring that they do not? Much of the previously cited literature examined how humans are biased and unreliable thinkers in comparison to simple algorithms, so we could all do better to emulate them. But is communication best when we try to be like formulae? Is science? If so, the human element is obsolete. Computer programs are very good (and getting better) at automating a lot of the repetitive and routine work necessary to research—not just writing styles but statistical analyses, database searches, and information consolidation to name but a few. That is to say an artificial intelligence program could today write a disturbingly large proportion of my thesis, not just those two paragraphs. It so follows that writing a technical paper myself will soon be the equivalent doing statistical calculations by hand today: worryingly error-prone and outdated. Most of the algorithms I have described mimic existing human processes in order to *replace* them; therefore, if we want humans to continue having an input in scientific research we need to work out what we—the cognitively more variable humans—can now offer to *complement* these processes. We need to expand our methods, not just emulate machines.

Moving humans away from machine thinking is more difficult than I would like to believe. It seems to me more and more that psychology and technology have been moving along the same road starting from different directions—and are about to meet. Psychology has been trying to convert human behaviour into understandable patterns and systems, using replicable routines and increasingly sophisticated methods of standardising both the activity of the observed and the observing. This research project is one example of that process, and the huge quantitative data banks held by IAPT are another. I confess to a certain amount of robotic behaviour myself, from reducing the participants to symptom counts, to defining 'good outcomes' as the difference between numbers, and conducting myself according to a combined 300 pages of protocol and style guidance. This is mine and psychology's inheritance from an Enlightenment style of thinking, viewing humans as organic machines on a quest for systematic empiricism, which we can still see entrenched in scientific thought today.

> "For me, the greatest achievement of Watson and Crick was to turn genetics from a branch of wet and squishy physiology into a branch of information technology, in the process slaying...the ghost of vitalism"
> (Dawkins, 2009, p. 226).

At exactly the same time, technologists have made every effort to make machines more anthropoidic (albeit less wet). In the 17th Century, while Descartes was conceiving of men as partially mechanical, he also postulated machines would one day speak and move (Descartes, 1637). After this was accomplished in the 20th Century, Turing proposed his 'Imitation Game' (Turing, 1950), which argues a computer can be said to exhibit human-level intelligence when it is mistaken for a person in conversation. In the early 21st Century, the challenge is arguably met (Veselov, 2014). The next stage in technological evolution involves moving away from controlled, replicable routines to a more unpredictable learning—teaching machines to think. So here we are, almost at the meeting point of man and machine. But where to go? If we keep going forward, we cover old ground.

14. Questions from a dystopian future

If we imagine a near future where machine learning is ubiquitous in psychological healthcare, yet functions relatively independently of its human overlords, we are left with some tantalising questions. For example, should an algorithm have a duty of care, in that it is designed to reasonably ensure harm does not occur to its users? Imagine that a person's online behaviour could be matched to profiles of psychosis so that they are 'diagnosed' weeks before they themselves are aware of symptoms. Should Google have a responsibility to diagnose and refer them, given otherwise the person will not access treatment for another 18 months (Hardy, Niendam & Loewy, 2017)? If someone's anorexia becomes worse from viewing online content, does the responsibility lie with Facebook? Some argue it should (Perrin & Woods, 2018), yet what does that mean for individual responsibility? Or that of their eating disorders care team? What does it mean for the therapeutic relationship when your therapist is a chat bot? What happens when (let us be honest, *when*) an algorithm 'learns' that constructs like Generalised Anxiety Disorder do not best fit the trillions of data nodes available to it and rewrites the Diagnostic and Statistical Manual? Our service user now no longer has anorexia, but 'Psypression—Variant 7', which will best respond to Individual Treatment Package #253.a.iii; unfortunately their local IAPT is still offering manualised high or low intensity treatment for the 'common' mental health diagnoses. How can we create an NHS responsive to that level of change? Are we humans negligent if we do not?

I have already mentioned some of the design issues elsewhere, such as the building in of biases or lack of service user data. One problem specifically concerns priorities: do you design for health or profit? As ethical clinicians, I am sure we would design to maximise psychological gains (as clinicians would define them), yet as commissioning realists I am equally sure we would design for cost minimisation. Whoever makes these algorithms will shape the future of healthcare, so should that be NHS psychologists or Microsoft? Since as a trainee psychologist I have not yet been offered a course on Basic Concepts in Machine Learning, I have a fairly reasonable idea of who will be assigning my future caseloads.

I am not convinced, however, that we need to accept a future where algorithms will do their work without us and we will simply marvel at their black boxes of thought until we become obsolete. I think we can offer something more than pale imitations of standardisation, which if done well might answer some of these questions before we get to the near dystopian future.

15. Putting the ‘soul’ back into psychology

I posit that now we do not have to disavow humanity in order to make great science any more; rather the opposite. We could leave protocol-driven work to the machines (they are more competent than us) and bring some personality to the room. I take the literature review as my example. I started it wanting to be as rigorous as possible in developing it, so followed a Cochrane Review protocol as strictly as I could—that being the ‘gold standard’ for literature reviews (and therefore ripe for automation). However, at the analysis section I faltered, as I did not know what data I would have, so could not predetermine what to do with it[5]. After starting the data collection I came across Pawson's (2004) method for realist syntheses, which offered solutions for understanding the data I was finding even though it sets itself up almost as a competitor with the Cochrane system (p. 14): flexibility against rigidity. In answering whether PCDSSs were effective, if I had fully stuck to my original rigid protocol I would have concluded with the highly respectable answer for a systematic literature review: sometimes[6].

At this point I would like to point out we now both have the prescient ability of knowing the conclusion of every similar question for a literature review that ever has or will be written. It is not exactly satisfying.

This is where humanity comes in. Pawson argued the usefulness of a review is not in establishing *if* something works but *when* it works, and this can be discovered using sources such as stakeholder discussions, personal experience and brain-storming, in addition to more traditional data. The 'when' is based on the people. For me, suddenly the most relevant items in academic papers were everything except the conclusion; every procedural difficulty, every moment of wry humour, every participant quote and acknowledgement that it did not work out how they expected, were all clues (some are given below in Box 1). The exceptions, deviations, and idiosyncrasies became the keys to understanding when and why things worked, not something to be controlled for. Synthesising the findings required an understanding of human behaviour, not only picking out themes and numbers. The resulting conclusions are so much richer and more useful because of this, and the process complements the more rigid database search that came before. It is, I hope, an example of how personality and protocols can work together to create respectable, auditable findings.

[5] This is not normally the case in Cochrane reviews, as eligible study outcomes should be more strictly defined, but I wanted to cast a large net

[6] Even the mobile monster game Pokémon GO improves mental health in certain people (Van Ameringen et al., 2017)

> Five of the six [participants] seemed to derive some pleasure when the machine gave the same recommendation as they did. One went so far as to exclaim, "Great!”
> (Benbenishty & Treistman, 1998, p. 201)

> He further stated that although the process of inputting the data is very healthy and positive for the decision maker (in that it forces the decision maker to review the information again), the recommendation given at the end is "worthless." To prove his point, this [participant] took on himself the challenge of finding a case to input that would stump the [PC]DSS - a task he accomplished.
> (Benbenishty & Treistman, 1998, p. 201)

> GPs expressed a general lack of familiarity with the subgrouping tool:
> GP17: I haven’t accessed your tool for about a year to be honest
> (Sanders, Foster, & Ong, 2011, p. 7)

> The [handouts] were apparently valued by the clinicians who began using the forms among patients in the [wrong] arm of the study. The resulting contamination would create a bias toward the null and may mean the study underestimated the effect of the[m].
> (Carroll, Biondich, Anand, Dugan, & Downs, 2013, p. 315)

Box 1: A selection of extracts from the reviewed literature highlighting some of the personal elements that informed the literature review findings

The appeal of this approach for me is personal, as it overlaps with many of the techniques developed in the technology sectors for optimising human-computer interaction that I hold in some esteem. Such tools include empathy mapping, experience journeys, visualisation and co-creation (see Liedtka, Oglive and Brozenske, 2014, for an overview), several of which can be seen in the case reports and service-related research project later on. These are not just typical qualitative data methods, as they also look at idiosyncrasies and unintended effects in order to understand and improve their systems. The difference between using these techniques to inform development and not is the difference between using a new smart phone without needing to read the manual and a two-day training course on your workplace's clinical notes programme. Several of the reviewed papers, including Buckingham et al. (2015) and Colombet et al. (2003), used these approaches in their algorithmic tool design, and they are also becoming more popular as levers for understanding and designing social change (for example see Stroh, 2015). Thus they can generate knowledge from an individual or object-based level and up to a social or service-based level.

16. Return of the hypotheses

In the final section of this reflective piece I would like to highlight some of the times more human-informed research has been possible in the research project, and times it has not, using the hypotheses developed at the end of the literature review for guidance.

16.1. Context

Many of the hypotheses included ideas on designing mental health algorithms around the situations they were to be deployed in, such as scoping existing procedures, involving stakeholders, and incorporating what was held to be important. The research has done this to a certain degree, for instance the data used to compare the different models is specific to the site that intends to use it, and clinicians and managers were frequently consulted on the feasibility of deploying the algorithm in practice. The algorithm itself is also an example of a conditional model of treatment, as it adapts to the individual. User-centred design processes that would examine context more explicitly, such as participant observation, were however missing from the research.

According to the guidelines from the literature review, chances of success could be increased by consulting service users for their views on computer-informed allocation, and having more information on IAPT allocation processes. For instance, the Services have now introduced a third allocation pathway for clients judged unlikely to benefit from standard high- or low-intensity therapy. The algorithm has not taken this into account, so cannot be used for a proportion of clients. This is a risk to uptake.

Interestingly, because the research looked at therapeutic outcomes at this particular site rather than using the established research base to assume knowledge, the algorithm suggests a model of care slightly antithetical to the IAPT system. IAPT is meant to be a ‘stepped care’ pathway, where clients are started on the lowest viable intensity of treatment and ‘stepped up’ to a higher intensity as necessary. The current research does not recommend this as both the MAP and ML algorithms suggested very few profiles would benefit from stepping up. Why this happened is unclear, but I find it an intriguing result.

16.2. Values

A lot has been written in the area of human-centred design about user needs analyses and requirement engineering to understand what it is people want (for example see Liedtka, 2017, for a description on co-designing residential care for autistic adults). As in psychological therapy, determining motivations for behaviour is not always straightforward.

For the research project, which was started before the conclusion of the literature review the emphasis was always on improving recovery rates, as that was believed to be the primary motivation of IAPT for using it. Yet on concluding the research I think a more important driver might be cost. This is partly because Hypothesis 2 argues outcomes are improved when resources are used more efficiently, which the MAP algorithm does. As it stands today, MAP would *not* improve recovery rates but would save an estimated 20% of the Services' therapy costs.

Taking the previous consultations at face value, it would be reasonable to assume the Services would not want to use the algorithm because cost was never mentioned and it does not address the primary value of recovery. Nevertheless, I think it likely (albeit a hunch) they would want to continue with deployment because cost *is* a motivational factor that was not identified. If the research was done again, I would spend more time developing an understanding of what is important to IAPT clinicians and managers in order to design the research more appropriately.

16.3. Symbiosis

Throughout this thesis, I have made much of how humans and machines need to complement each other in order to work effectively. The algorithm does this by providing 'recommendations' for therapeutic intensity, so the clinician will always make the final decision. It this way it fulfils the criteria of working with the user and allowing the clinician to exercise discretion, and makes it more likely to be adopted. However, this interdependence is largely superficial and was not a consideration in the research project. If the algorithm was implemented more widely I have some concerns of how the mutually-symbiotic relationship would continue.

First, the algorithm was built in a way that functions independently of the clinician, in that a recommendation for treatment will always be provided whether the clinician makes a decision or not. Second, the present research argues comparable effectiveness between totally human allocation and totally machine allocation, suggesting the algorithm could theoretically entirely replace its biological counters without detriment. Third, the barriers to doing this in practice are not so great, especially as many IAPT services already request service users complete the questionnaire the algorithm uses online. A few extra screening questions, such as 'have you had IAPT treatment before?' or 'do you need emergency support?', could—again theoretically—remove the need for a clinician in the majority of cases. Fourth, the benefits of doing this come under the heading of 'more efficient use of resources' (Hypothesis 2), since making the assessment process online and algorithm-based would save time for clients, money on clinicians, and space for assessment centres. If the cost of an assessment is £99 (based on Radhakrishnan et al., 2013) and IAPT sees 1,500,000 people annually (Nuffield Trust, 2018), that is a potential unadjusted 'saving' just shy of £150 million, or 8% of the NHS deficit (Dunn, Mckenna, & Murray, 2016) without taking into account further savings from the therapy allocated itself. So there is certainly the means and motivation to end the symbiotic relationship between algorithm and clinician. Developing a better understanding of how that symbiosis can be strengthened through user-centred investigations would be important if this outcome were to be avoided.

16.4. Trust

I have tried to be transparent in articulating the processes behind the algorithm, which I hope will foster a sense of trust from understanding in anyone interested in using the algorithm. In Hypothesis 4 I posed several questions an algorithm needs to demonstrably answer to engender trust from its users:

1. How well does it support decisions for complex clients?
2. How well does it take contextual factors (organisational policy, resource availability, targets, etc.) into account?
3. Does it do both of these better than the user?

In answer to question 1, I think the algorithm does this reasonably well by differentiating the different profiles (profiles 5, 6, and 7 for example represent relatively complex cases with higher levels of clinical and social issues) and allowing the clinician to make their own decision. For question 2, again I think this is done reasonably well but does include some limitations, particularly regarding the third pathway as discussed earlier. For question 3, the research project suggests it performs equally well as a clinician, however the critical factor is how visible this transparency will be (forgive the pun). For instance, it is not reasonable to expect most/any IAPT clinicians to read this thesis, regardless of how clear the conclusions are. A bite-size version of the relevant findings might be useful in this case.

One major unknown is whether IAPT service users will trust the tool, as there was little from the literature review to answer that question and no additional consultation with them. It is also unclear whether they would want answers to the same questions as clinicians and how best to show these. If I did similar research in the future, I would include a service user consultation group early on.

17. Conclusion

We live in interesting times where algorithms can not only mimic humans but surpass them in certain ways. They reside in our homes, sit in our phones, and stare back at us from our work screens. Yet it is not their advancement that I have written this last reflection on, but ours. Empirical science has traditionally been about ‘rationalist’ thought, bias elimination and universal truths, for which we humans are woefully ill-prepared to be faithful too. Sticking to a rigid set of rules can rob us of our humanity in ways that are no longer relevant and currently prevent us doing better research. I personally look forward to creating more human-informed studies in the future, and including people's values and contexts with more systematic evidence sources. I only hope you trust it.

18. References

Benbenishty, R., & Treistman, R. (1998). The development and evaluation of a hybrid decision support system for clinical decision making: The case of discharge from the military. *Social Work Research*, *22*(4), 195–204.

Buckingham, C. D., Adams, A., Vail, L., Kumar, A., Ahmed, A., Whelan, A., & Karasouli, E. (2015). Integrating service user and practitioner expertise within a web-based system for collaborative mental-health risk and safety management. *Patient Education and Counseling*, *98*(10), 1189–1196. http://doi.org/10.1016/j.pec.2015.08.018

Carroll, A. E., Biondich, P., Anand, V., Dugan, T. M., & Downs, S. M. (2013). A randomized controlled trial of screening for maternal depression with a clinical decision support system. *Journal of American Medical Informatics Association*, *20*(2), 311–316. http://doi.org/10.1136/amiajnl-2011-000682

Colombet, I., Dart, T., Leneveut, L., Zunino, S., Menard, J., & Chatellier, G. (2003). A computer decision aid for medical prevention: a pilot qualitative study of the Personalized Estimate of Risks (EsPeR) system. *BMC Medical Informatics and Decision Making*, *3*(13). http://doi.org/10.1186/1472-6947-3-13

Dunn, P., Mckenna, H., & Murray, R. (2016). Deficits in the NHS 2016. *The King's Fund*, 36. Retrieved from http://www.kingsfund.org.uk/sites/files/kf/field/field_publication_file/Deficits_in_the_NHS_Kings_Fund_July_2016_1.pdf

Hardy, K. V., Niendam, T. A., & Loewy, R. (2017). Measuring the Duration of Untreated Psychosis within First Episode Psychosis Coordinated Speciality Care. National Association of State Mental Health Program Directors. Available from www.nasmhpd.org

Pawson, R., Greenhalgh, T., Harvey, G., & Walshe, K. (2004). Realist synthesis - an introduction. *ESRC Research Methods Programme*, (January 2004), 1–46. Retrieved from http://discovery.ucl.ac.uk/180102/

Perrin, W., and Woods, L. (2018). Reducing Harm In Social Media Through A Duty Of Care. Carnegie UK Trust. Available from http://blogs.lse.ac.uk/

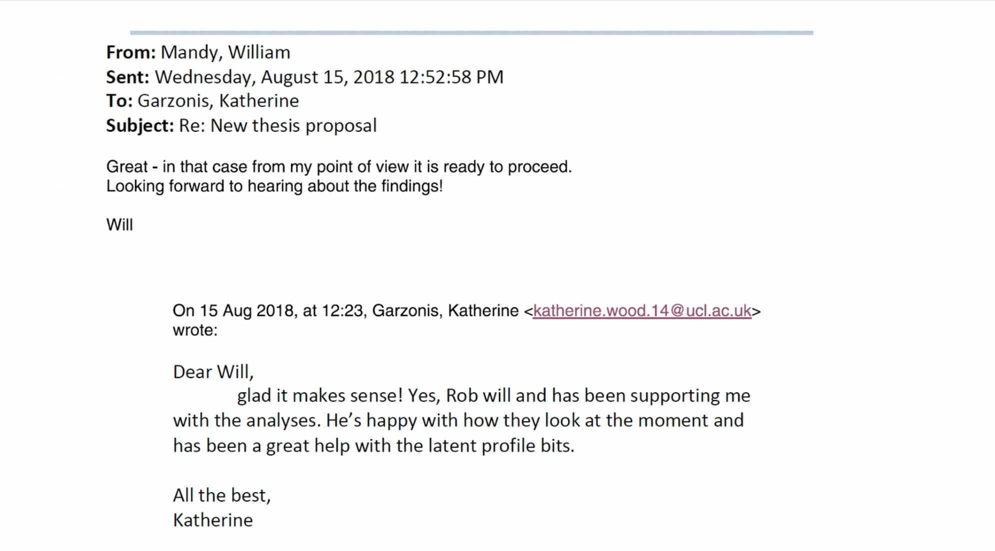
Radhakrishnan, M., Hammond, G., Jones, P. B., Watson, A., McMillan-Shields, F., & Lafortune, L. (2013). Cost of Improving Access to Psychological Therapies (IAPT) programme: An analysis of cost of session, treatment and recovery in selected Primary Care Trusts in the East of England region. *Behaviour Research and Therapy*. http://doi.org/10.1016/j.brat.2012.10.001

Sanders, T., Foster, N. E., & Ong, B. N. (2011). Perceptions of general practitioners towards the use of a new system for treating back pain: a qualitative interview study. *BMC Medicine*, *9*(49).

Turing, A. M. (1950). Computing machinery and intelligence-AM Turing. *Mind*, *49*, 433–460.

APPENDICES

Appendix A : Email Exchange Denoting Ethical Approval



From: Mandy, William
Sent: Wednesday, August 15, 2018 12:52:58 PM
To: Garzonis, Katherine
Subject: Re: New thesis proposal

Great - in that case from my point of view it is ready to proceed.
Looking forward to hearing about the findings!

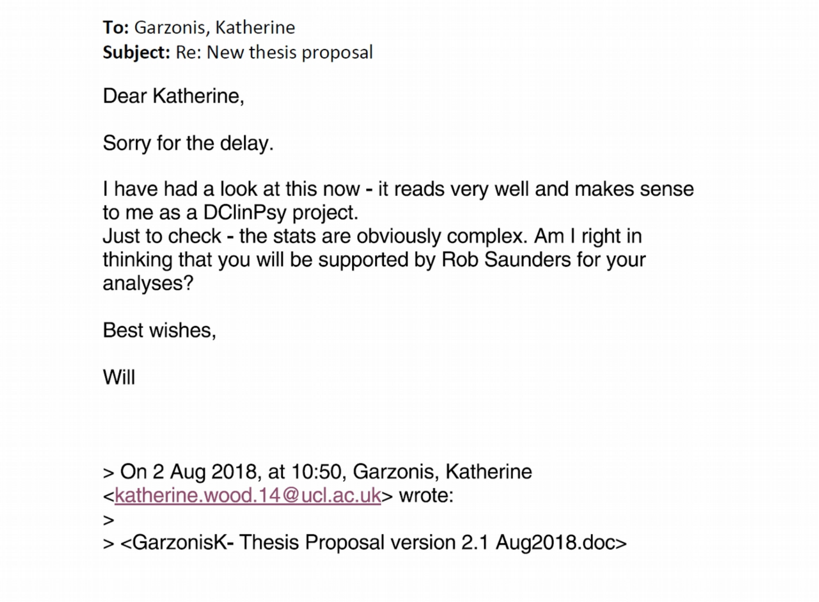
Will

On 15 Aug 2018, at 12:23, Garzonis, Katherine <katherine.wood.14@ucl.ac.uk> wrote:

Dear Will,

glad it makes sense! Yes, Rob will and has been supporting me with the analyses. He’s happy with how they look at the moment and has been a great help with the latent profile bits.

All the best,
Katherine



To: Garzonis, Katherine
Subject: Re: New thesis proposal

Dear Katherine,

Sorry for the delay.

I have had a look at this now - it reads very well and makes sense to me as a DClinPsy project.
Just to check - the stats are obviously complex. Am I right in thinking that you will be supported by Rob Saunders for your analyses?

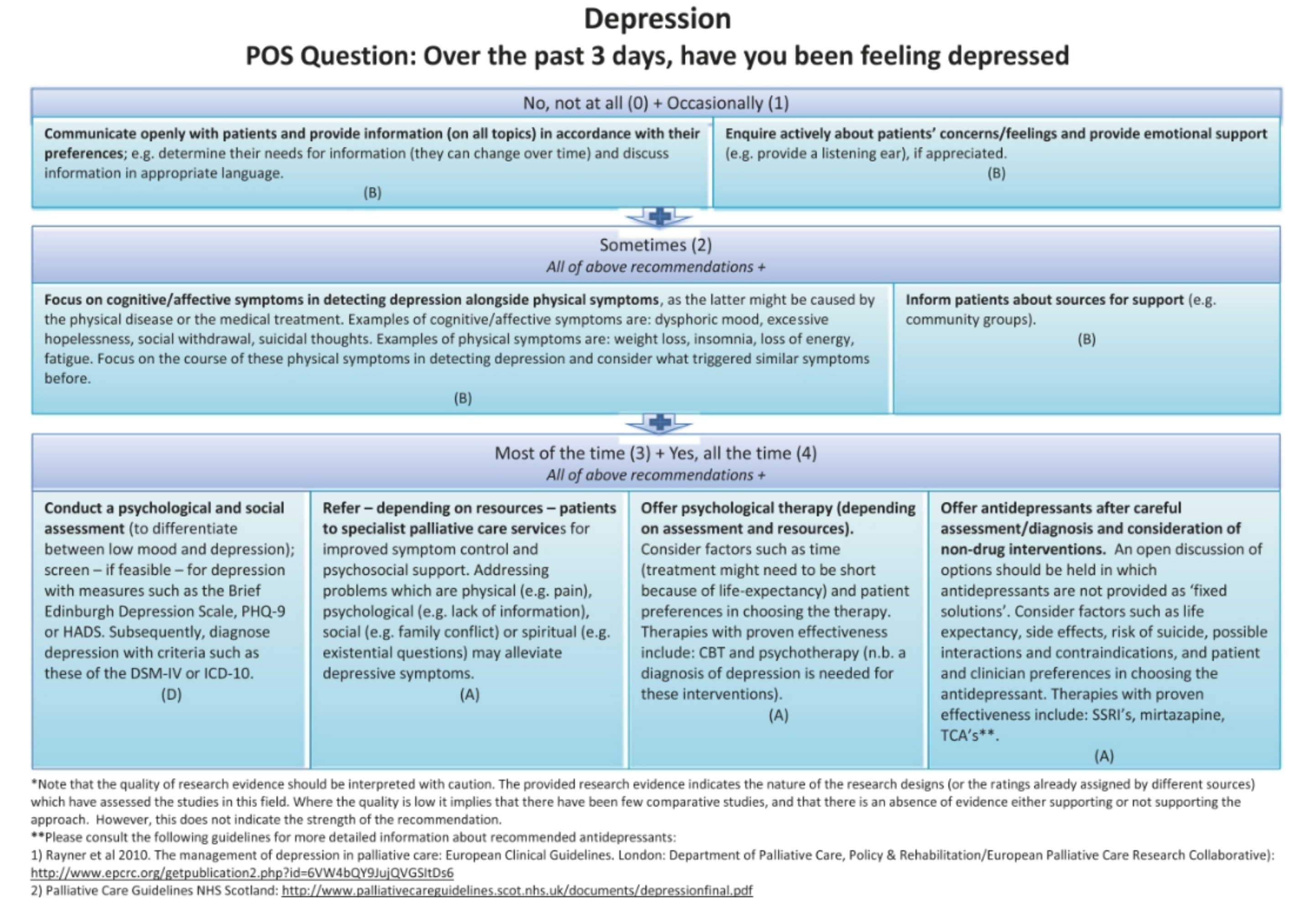
Best wishes,

Will

> On 2 Aug 2018, at 10:50, Garzonis, Katherine <katherine.wood.14@ucl.ac.uk> wrote:
>
> <GarzonisK- Thesis Proposal version 2.1 Aug2018.doc>

Appendix B : An example of a Predictive Clinical Decision Support System

The diagram below is taken from van Vliet et al. (2015, p. 12), who designed a system to aid palliative care staff identify several patient needs. It looks at symptoms of depression. Given the client response to the screener question at the top, staff are recommended to proceed with particular actions. Letters in brackets refer to the strength of evidence for that action.



Depression

POS Question: Over the past 3 days, have you been feeling depressed

No, not at all (0) + Occasionally (1)

Communicate openly with patients and provide information (on all topics) in accordance with their preferences; e.g. determine their needs for information (they can change over time) and discuss information in appropriate language.
(B)

Enquire actively about patients' concerns/feelings and provide emotional support (e.g. provide a listening ear), if appreciated.
(B)

Sometimes (2)
All of above recommendations +

Focus on cognitive/affective symptoms in detecting depression alongside physical symptoms, as the latter might be caused by the physical disease or the medical treatment. Examples of cognitive/affective symptoms are: dysphoric mood, excessive hopelessness, social withdrawal, suicidal thoughts. Examples of physical symptoms are: weight loss, insomnia, loss of energy, fatigue. Focus on the course of these physical symptoms in detecting depression and consider what triggered similar symptoms before.
(B)

Inform patients about sources for support (e.g. community groups).
(B)

Most of the time (3) + Yes, all the time (4)
All of above recommendations +

Conduct a psychological and social assessment (to differentiate between low mood and depression); screen – if feasible – for depression with measures such as the Brief Edinburgh Depression Scale, PHQ-9 or HADS. Subsequently, diagnose depression with criteria such as these of the DSM-IV or ICD-10.
(D)

Refer – depending on resources – patients to specialist palliative care services for improved symptom control and psychosocial support. Addressing problems which are physical (e.g. pain), psychological (e.g. lack of information), social (e.g. family conflict) or spiritual (e.g. existential questions) may alleviate depressive symptoms.
(A)

Offer psychological therapy (depending on assessment and resources). Consider factors such as time (treatment might need to be short because of life-expectancy) and patient preferences in choosing the therapy. Therapies with proven effectiveness include: CBT and psychotherapy (n.b. a diagnosis of depression is needed for these interventions).
(A)

Offer antidepressants after careful assessment/diagnosis and consideration of non-drug interventions. An open discussion of options should be held in which antidepressants are not provided as 'fixed solutions'. Consider factors such as life expectancy, side effects, risk of suicide, possible interactions and contraindications, and patient and clinician preferences in choosing the antidepressant. Therapies with proven effectiveness include: SSRI's, mirtazapine, TCA's**.
(A)

*Note that the quality of research evidence should be interpreted with caution. The provided research evidence indicates the nature of the research designs (or the ratings already assigned by different sources) which have assessed the studies in this field. Where the quality is low it implies that there have been few comparative studies, and that there is an absence of evidence either supporting or not supporting the approach. However, this does not indicate the strength of the recommendation.

**Please consult the following guidelines for more detailed information about recommended antidepressants:

1) Rayner et al 2010. The management of depression in palliative care: European Clinical Guidelines. London: Department of Palliative Care, Policy & Rehabilitation/European Palliative Care Research Collaborative): http://www.epcrc.org/getpublication2.php?id=6VW4bQY9JujQVGSltDs6

2) Palliative Care Guidelines NHS Scotland: http://www.palliativecareguidelines.scot.nhs.uk/documents/depressionfinal.pdf

Fig. 8 Final decision diagrams. Legends: POS score decision diagrams format 2. Depression

Appendix C : Literature Review Search Protocol

C.1 Change Record

Version 1, 18 Feb 2017

Version 2, 28 Mar 20187: clarification of search strategy for individual databases added. Unsearchable databases removed from search list. Clarification of exclusion and inclusion criteria added.

C.2 Background

C.2.1 Why there is a need for a study on this topic

Clinical decision making is prone to human errors and biases (Elstein and Schwarz, 2002). Mechanical predictions have been found to perform as well as or better than human clinical judgements (Grove et al., 2000), and therefore could be used to improve clinical decision making. Studies exist that trial the effectiveness of such tools in psychological practice, yet rarely are these analysed together.

The current study is interested in the use of predictive tools; that is, rule-based systems involving machine calculation that can be used to prospectively determine who can benefit from a particular intervention. This can include recommendations for therapy based on analysis of wellbeing scores, suggestions for further assessment after a new diagnosis, and other such proactive measures. To the author's knowledge, a systematic literature review of prospective clinical decision support systems (PCDSS) in mental health settings has not so far been attempted.

C.2.2 Main research question

What is the evidence for the use of prospective tools in clinical decision making to improve mental health treatment outcomes in clients?

C.2.3 Additional research questions that will be addressed

1. What prospective clinical decision support systems (PCDSS) are effective in mental health?
2. What are the risks associated with PCDSSs in mental health settings?

C.3 Search Strategy

C.3.1 Basic strategy

Automated search of electronic databases will be used to locate the majority of the articles, as this will speed up the rate of identification. Added to these results will be any relevant documents previously identified as part of background reading, which will be subject to the same inclusion and exclusion criteria. This will ensure a greater quality of materials are analysed in a timely way.

Only studies that use an experimental or quasi-experimental design to test the clinical value of a PCDSS will be included as evidence for research question 1. If criteria are not met, data pertaining to question 2 will still be used. Research methodology will not be specified in the search string, as this will minimise the number of appropriate studies missed through inadequate filters (Gorecki et al., 2010).

C.3.2 Search terms

A prototype search string was created using key terms present in known studies and structured according to the research design implied by the review questions (see Table 1 below). Terms were also cross-referenced with index terms or thesaurus capabilities present in some research databases, such as PubMed, in order to generate as many related and relevant terms as possible. The resulting string was tested in databases know to contain previously identified studies to determine whether it was sensitive enough to detect these. If too many spurious results were obtained, the string would be adapted to minimise this. Subsequent search strings were revised according to this method.

Table C1

Search terms arranged by substring type, with their logical operators

Substring	Search Terms
Area of interest	AND (“decision making” OR “decision-making” OR “clinical judgement” OR “clinical decision” OR predict)
Intervention	AND (tool OR “decision support system” OR “decision rules” OR algorithm)
Outcomes	AND (“mental health” OR “mental illness” OR wellbeing OR depression OR anxiety OR admission OR discharge OR referral OR “quality of life”)
Population	AND (psychologists OR professionals OR clinicians)

Search string version 1

- Additional equivalent terms added for population substring based on previously identified studies: OR practitioner* OR provider* OR physician*
- Additional outcome criteria added to substring to expand number of relevant studies returned: OR “treatment response” OR “response to treatment”
- ‘System’ removed from intervention substring, which would otherwise exclude a previously identified relevant study. This expands the potential number of studies identified.
- Additional intervention terms added to expand number of relevant studies returned based on previously identified studies: OR aid OR “care suggestions” OR “treatment advice”
- Area of interest expanded to include search term “care suggestions” based on previously identified studies.

Search: Title and Abstract

AND (“decision making” OR “decision-making” OR “clinical judgement” OR “clinical decision” OR predict OR “care suggestions”)

AND (tool OR “decision support” OR “decision rules” OR algorithm OR aid OR “care suggestions” OR “treatment advice”)

AND (“mental health” OR “mental illness” OR wellbeing OR depression OR anxiety OR admission OR discharge OR referral OR “quality of life” OR “treatment response” OR “response to treatment”)

AND (psychologist* OR professional* OR clinician* OR practitioner* OR provider* OR physician*)

Search string version 2

- Added an exemption for studies on shared decision-making to remove spurious results: NOT “shared decision”
- Added exemption for systematic reviews in order to remove unnecessary results: NOT “systematic literature” OR “systematic review”

Search: Title and Abstract

AND (“decision making” OR “decision-making” OR “clinical judgement” OR “clinical decision” OR predict OR “care suggestions” OR “care process*”)

AND (tool OR “decision support” OR “decision rules” OR algorithm OR aid OR “care suggestions” OR “treatment advice”)

AND (“mental health” OR “mental illness” OR wellbeing OR depression OR anxiety OR admission OR discharge OR referral OR “quality of life” OR “treatment response” OR “response to treatment”)

AND (psychologist* OR professional* OR clinician* OR practitioner* OR provider* OR physician*)

NOT “shared decision”

Search: Title
NOT (“systematic literature” OR “systematic review”)

Search string version 3

- Added full text filter to ensure better conformity with in/exclusion criteria. Note that the filter is not for ‘free full text availability’.
- The use of index terms to filter for human participant studies was considered, as per the inclusion criteria. However, it was found using this excluded potentially relevant human studies that had not been associate with the ‘human’ index term on some databases. Note that non-human index terms, such as the category ‘rats’, can be used to exclude as the rate of false positives is sufficiently low. If it is not possible to exclude non-human studies directly through a database (as some will only filter for terms, not filter out), the same effect can be achieved by using the search terms without a species filter (Search String [A]), re-running the search with animal filters (Search String [B]), and then building a third search using the two saved terms (Search String [A-B]).

Search: Title and Abstract
AND (“decision making” OR “decision-making” OR “clinical judgement” OR “clinical decision” OR predict OR “care suggestions” OR “care process*”)

AND (tool OR “decision support” OR “decision rules” OR algorithm OR aid OR “care suggestions” OR “treatment advice”)

AND (“mental health” OR “mental illness” OR wellbeing OR depression OR anxiety OR admission OR discharge OR referral OR “quality of life” OR “treatment response” OR “response to treatment”)

AND (psychologist* OR professional* OR clinician* OR practitioner* OR provider* OR physician*)

NOT “shared decision”

Search: Title
NOT (“systematic literature” OR “systematic review”)

Filter: Full text available

Filter: Remove results attached to non-human index terms

Search string version 4

- Wildcards removed after it was realised they were not functioning as they should on some databases. They have been replaced with their intended plural equivalents (e.g. ‘professional OR professionals’)

Search: Title and Abstract

AND (“decision making” OR “decision-making” OR “clinical judgement” OR “clinical decision” OR predict OR “care suggestions” OR “care process” OR “care processes”)

AND (tool OR “decision support” OR “decision rules” OR algorithm OR aid OR “care suggestions” OR “treatment advice”)

AND (“mental health” OR “mental illness” OR wellbeing OR depression OR anxiety OR admission OR discharge OR referral OR “quality of life” OR “treatment response” OR “response to treatment”)

AND (psychologist OR psychologists OR professional OR professionals OR clinician OR clinicians OR practitioner OR practitioners OR provider OR providers OR physician OR physicians)

NOT “shared decision”

Search: Title

NOT (“systematic literature” OR “systematic review”)

Filter: Full text available

Filter: Remove results attached to non-human index terms

Search Terms to use (Version 4, 19 Feb 2017):

Search: Title and Abstract

AND (“decision making” OR “decision-making” OR “clinical judgement” OR “clinical decision” OR predict OR “care suggestions” OR “care process” OR “care processes”)

AND (tool OR “decision support” OR “decision rules” OR algorithm OR aid OR “care suggestions” OR “treatment advice”)

AND (“mental health” OR “mental illness” OR wellbeing OR depression OR anxiety OR admission OR discharge OR referral OR “quality of life” OR “treatment response” OR “response to treatment”)

AND (psychologist OR psychologists OR professional OR professionals OR clinician OR clinicians OR practitioner OR practitioners OR provider OR providers OR physician OR physicians)

NOT “shared decision”

Search: Title

NOT (“systematic literature” OR “systematic review”)

Filter: Full text available

Filter: Remove results attached to non-human index terms

If it is not possible to search a part of the document specifically, such as the title, with a given substring, then the following substitutions may be made:

Table C2
Acceptable substitute search areas for particular substrings, arranged by order of preference

Substring	Ideal search area	Substitute search area(s)
AND ("decision making" OR "decision-making"…	Title and Abstract	Abstract only Full text Everywhere
AND (tool OR "decision support"...	Title and Abstract	Abstract only Full text Everywhere
AND ("mental health" OR "mental illness"...	Title and Abstract	Abstract only Full text Everywhere
AND (psychologist OR psychologists...	Title and Abstract	Abstract only Full text Everywhere
NOT "shared decision"	Title and Abstract	Title only Abstract only Remove from search string
NOT ("systematic literature" OR "systematic review")	Title	Remove from search string

Note the last two substrings can be removed from the search terms without having a large negative impact on the search results, although it will create more spurious hits.

If during the course of a search no appropriate papers are discovered in the first 100 results, then the search may be assumed to be ineffectual and abandoned at that point.

C.3.3 Databases to be searched

A list of journal databases was compiled using the UCL Library catalogue. This listed 44 databases related to psychology, from which were excluded topics not relevant to the current search. These included databases unlikely to contain original studies (such as Cochrane's CDSR, which only indexes reviews), or areas not directly related to the current topic of interest (such as EMBASE, which focuses on pharmacological interventions, and is otherwise covered through MEDLINE and PubMed searches).

The final search list of 27 databases is as follows:

- Annual Reviews
- APPI Journals (Psychiatry Online)
- ASSIA Applied Social Sciences Index and Abstracts
- CINAHL Plus
- COPAC
- HAPI (Health and Psychosocial Instruments)
- IBSS (International Bibliography of the Social Sciences)
- IngentaConnect
- JISC Journal Archives
- Journals@Ovid
- JSOTR
- MEDLINE (Ovid)
- Nature Journals
- PEP (Psychoanalytic Electronic Publishing)
- PILOTS (Published International Literature On Traumatic Stress)
- ProQuest Central
- Psyc -ARTICLES, -EXTRA, -INFO, and -TESTS
- Pubget
- PubMed
- Science Citation Index Expanded
- ScienceDirect (Elsevier)
- SCOPUS
- University of London Research Library Services
- Wiley Online Library

C.3.4 Manual searches

None.

C.3.5 Time period to be covered

No time period specified.

C.3.6 Ancillary search procedures

Identified studies that do not meet inclusion criteria but are judged to be highly likely to be included if there is a follow-up study will be checked against citation lists. For instance, a validation study of a bipolar treatment decision tree would not be included in the final set of papers because it is not specifically used in the study by clinicians to influence a decision. However, if the paper identifies itself as a feasibility study for a wider research initiative that intends to do just this, and the original paper was published some years ago, it is reasonable to expect a qualifying follow-on study to exist. In such a case, the citation list for the original study will be eye-balled for further qualifying research. Any studies identified in this way will be subject to the same criteria and general process as studies found through the main search method.

C.3.7 How the search process is to be evaluated

- A senior researcher will be involved in the search strategy and study evaluation to test for inter-rated reliability
- Comprehensive search methods used to locate studies, such as multiple trials of search terms
- Thorough search of appropriate databases
- Potentially important sources explored through ancillary search procedures
- Study selection criteria determined before search is initiated
- Validity of studies assessed appropriately, and criteria reported
- Review methods clearly reported

C.4 Selection Criteria

C.4.1 Inclusion criteria for primary studies

- Quantitative experimental or quasi-experimental research design (criteria applies only to studies answering question 1; other designs may be included for question 2)

- Full text is available, either freely online or via the UCL Library services
- PCDSS is rule-based (personalised):
  - i.e. an automatic computation is carried out based on client/patient data by the PCDSS, and not only manually by a human agent (or may be feasibly carried out by a machine, such as scoring a simple likert-based questionnaire)
  - for instance, an automatic recommendation of a therapy based on (human or machine) input and (machine) analysis of depression and anxiety scores would be included; a human screening for people eligible to receive automatic reminders (with no further PCDSS calculation involved) would not be included
  - a paper-based decision-tree may be treated as an automatic computation for the purposes of this review, as the recommendation of decision based on a pathway of criteria practically functions in the same way as a computer algorithm
- Intervention could feasibly be used appropriately by psychologists in a mental healthcare setting:
  - this broadly includes all interventions designed to alter thinking or behaviour linked to mental health outcomes
  - for instance this would include tools involved with a reduction in smoking (which is linked to mental health outcomes {REF} and is considered a general NHS initiative [REF]), but would not include recommendations to initiate a cancer screening, which is a task only carried out by physical health professionals
- At least one reported outcome is relevant to mental healthcare settings:
  - e.g. reduction in anxiety symptoms, improved quality of life, etc.

  - improving treatment adherence for physical health problems such as diabetes may be included, as this is a function carried out by many health psychologists. However, if the intervention itself is not appropriate to a psychologist, e.g. education of how to manage diabetes with no inclusion of psychological factors, the study should be excluded
  - if the only reported outcome is how well the tool matches decisions made by professionals (which is not directly relevant to a mental health care setting), without the professionals using the tool, the study will be excluded
- PCDSS is intended to be used by clinicians to improve their decision-making, and is used in this context in the study
- PCDSS produces a decision/recommendation for action

C.5 Exclusion criteria

- Studies with only non-mental health outcomes, such as orthopaedic surgery recovery rates:
  - where it is unclear whether an outcome is relevant, they may be considered so if they are included as part of Public Health England (2014) Priorities, such as change in weight or alcohol dependency
- Systematic literature reviews are excluded from the analysis
- Research that is not prospective
- No appropriate comparison data, e.g. no pseudo-/control group (for experimental studies answering question 1 only)
- Full text not available (does not refer to Free Full Text availability)

- Written in a language other than English (or language not understood by the reviewer)
- Books (as unlikely to contain studies not available elsewhere, and would be inappropriately time-consuming to identify within the text if they did)
- Non-human subjects
- A non-PCDSS tool is used:
    - A decision-support system can be distinguished from, for instance, a diagnostic tool in that specific suggestions for action are created by the tool following the input of data. A diagnostic tool will only produce a label, not a recommendation for behaviour by the clinician.
- Decision utility not assessed
- The study does not examine at least one specific PCDSS
- PCDSS is intended to support the decision-making process of someone other than the clinician, e.g. the client
- The use of the PCDSS to inform decision-making is not studied:
    - e.g. the development and validation of a tool that is intended inform decision-making, but is not used for this purpose in the study in question

C.5.1 How selection will be undertaken (roles of analyst)

Table C3

Steps in the article selection and exclusion process

Step	Article Selection Process	Exclusions
1	Articles obtained through search terms	
	Articles obtained identified through other sources	
2		Exclude duplicate entries
		Exclude articles not relevant to the topic based on their title (or where title is ambiguous, cursory reading of abstract)
3	Abstracts retrieved for further evaluation	
4		Exclude those not meeting criteria
5	Full text asked for more detailed evaluation	
6		Exclude those not obtainable as full text
7	Full text obtained for more detailed evaluation	
8		Exclude those not meeting criteria
9	Follow-up studies on promising research that did not meet criteria in Step 7 examined	
10		Exclude follow-up studies not meeting criteria
11	Articles finally included and assessed for quality (qualitative and quantitative studies assessed with respective check-lists)	

C.5.2 How agreement among analysts will be evaluated

See 5. c) below.

C.5.3 Resolving differences between analysts

The analysts will attempt to reach an agreement by justifying their decisions with reference to the protocol criteria, or explain their reasons for wanting to change the protocol. In the majority of cases, resolution should be achieved through these two methods. If there is still no clear agreement, the decision will pass to an additional experienced researcher—who not directly responsible for conducting the analysis—for arbitration. In the event of a decision to alter the protocol, all searches will be re-run following the most recent version.

C.6 Study Quality Assessment

C.6.1 Quality checklists

The Mixed Methods Appraisal Tool (Pluye et al., 2011).

C.6.2 How the checklist will be evaluated

Study quality will be evaluated using the MMAT toolkit (Pluye et al., 2011). This has been chosen for its flexibility, as the same tool can be used on several different research designs. This can aid in the consistency of quality rating between studies, rather than for example using different tools for randomised control trials, quantitative, and mixed designs. The tool also produces a quality star rating, unlike some other frameworks, which may be useful as a general indicator of quality. It is acknowledged there are limitations to using such ratings for qualitative criteria: for example it would give equal weighting to a both representative sample and a high response rate, the relative impact of either of which is debatable. However, by using the ratings as an indication of quality, rather than an absolute value, high and low quality studies can be identified more easily than with some other measures.

See Appendices for full MMAT criteria.

C.6.3 How reliability of the data extraction method will be evaluated

Agreement will be assessed by having two researchers (the research author and a senior researcher unfamiliar with the study) extract data from the same database, and evaluate the studies for quality, as above. Their results will be compared for any significant deviation. Deviations will be examined and discussed where they arise to ensure both researchers follow the same methods. Changes to the protocol (e.g. further clarification) will be made where appropriate. This precess will be repeated with additional databases until consensus of method and results is achieved. This exercise will be carried out before the researcher author examines other databases, to help prevent errors or re-analysis in later searches.

C.6.4 How differences between data extractors will be resolved

As with 4)e.

C.6.5 Procedures to use for applying the checklists

See diagram in Appendices.

C.7 Data Extraction

C.7.1 Data extraction form

Researchers will fill out the following table for each paper selected for the review:

Table C4
Data extraction table

Study Details	**Population/setting**	**Delivery of intervention**	**Outcomes measures**	**Outcome statistics**	**Additional comments**
Author Year Design MMAT score	Intervention target (psychologists, nurses, etc.) Client characteristics (cardiac patients, anxiety diagnosis, etc.) Sample size	How was the intervention used? (e.g. integrated into existing systems, stand-alone programme, etc.)	Relevant measures Follow-up period	Statistical tests and results	Any problems with the analysis (e.g. insufficient statistical reporting, major biases, etc.)

C.7.2 Validation of the data extraction process

As with previous validation procedure, where two researchers undertake the same task, and a comparison on their respective results made for consistency.

C.8 Synthesis

C.8.1 Form of analysis to be used

Analysis will be separated by studies answering research question 1 (pseudo-/controlled experimental studies) and question 2 (all methodologies). Meta-analyses for question 1 and a mixed-method synthesis for question 2 will be attempted if appropriate data is obtained. Due to the high level of heterogeneity between study outcomes, a narrative account of the data is likely.

C.8.2 Assessing threats to validity

These will be noted and taken into account in the final analysis.

C.9 Study Limitations

C.9.1 Residual validity issues including potential conflicts of interest

These will be noted and taken into account in the final analysis.

C.10 Reporting

C.10.1 Target audience, relationship to other studies, planned publications, authors of the publications

The target audience is primarily the markers for the DClinPsy thesis. The literature review will therefore have to conform to standards laid out in the UCL DClinPsy Handbook. As the review will act as the introduction to the main research project, it is important it is reported in a way that logically precedes the main study, and can be clearly and conceptually linked to the work of Saunders et al. (2016), on which the study is based. The topic of literature review directly addresses these last two points, as Saunders is concerned with the development of a PCDSS, which is prospectively tested by the research project. The literature review can thus inform the usefulness (or not) and potential risks of such an undertaking.

It is hoped the literature review may be published at a later date as a stand-alone piece, where the target audience becomes academic psychologists. The authors for this are to be determined, depending on the level of additional input required for the review to be made publishable.

C.11 References

Elstein, A. S., & Schwarz, A. (2002). Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *Biomedical Journal,* 324, 729-732.

Gorecki, C. A., Brown, J. M., Briggs, M., & Nixon, J. (2010). Evaluation of five search strategies in retrieving qualitative patient-reported electronic data on the impact of pressure ulcers on quality of life. *Journal of Advanced Nursing, 66*(3), p. 645-52.

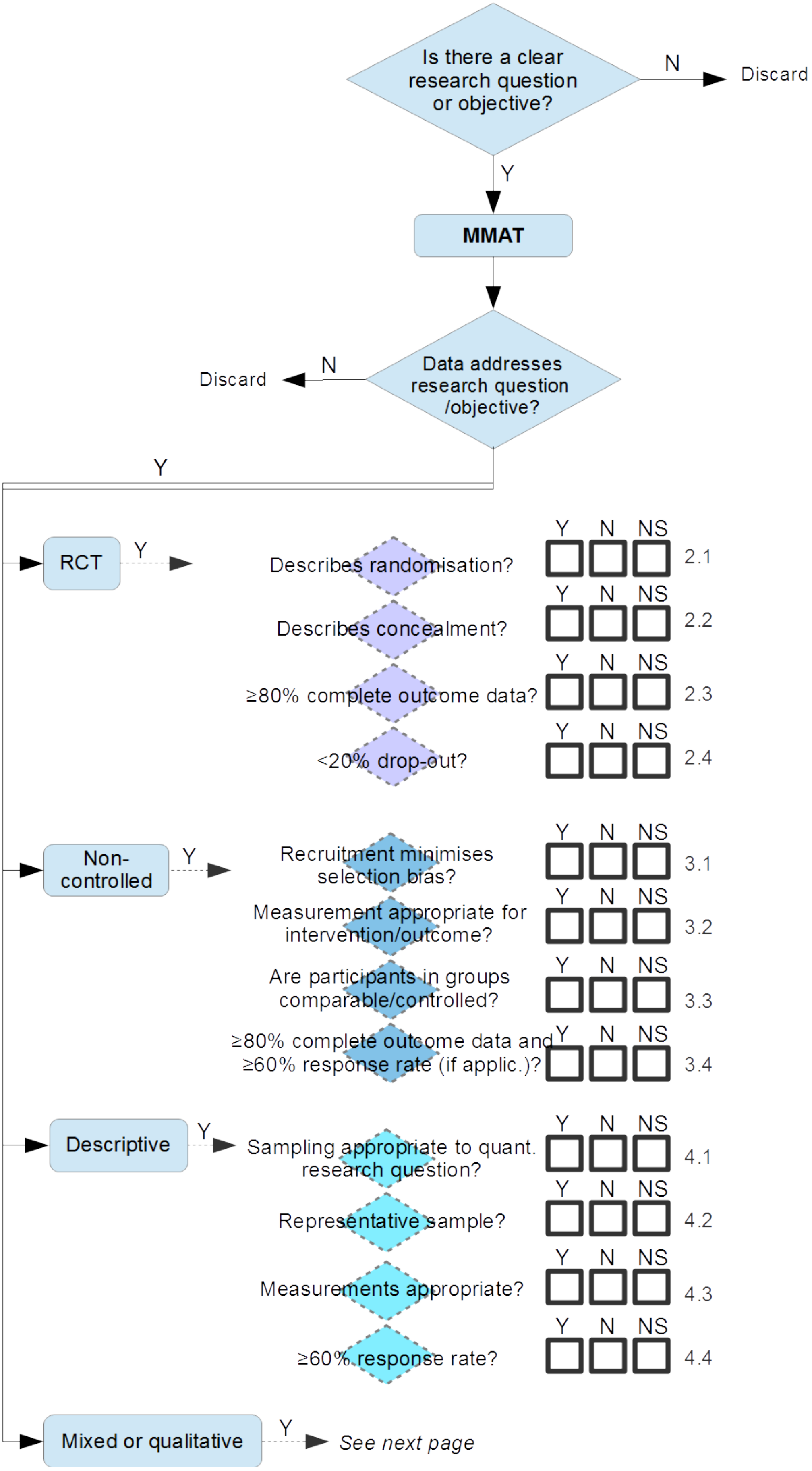
Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical Versus Mechanical Prediction: A Meta-Analysis. *Psychological Assessment, 12*(1), 19-30.

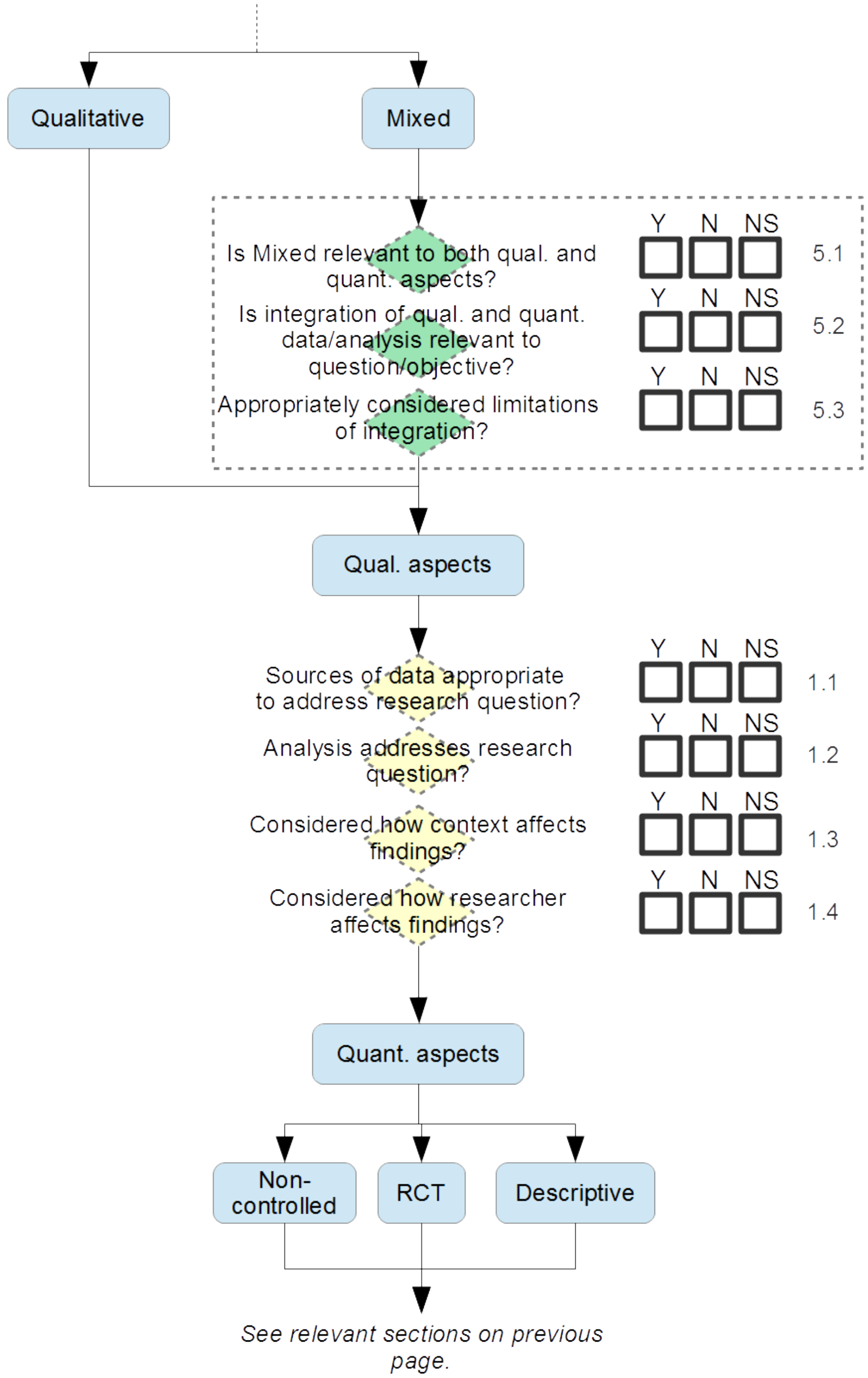
Pluye, P., Robert, E., Cargo, M., Bartlett, G., O'Cathain, A., Griffiths, F., Boardman, F., Gagnon, M.P., & Rousseau, M.C. (2011). *Proposal: A mixed methods appraisal tool for systematic mixed studies reviews.* Retrieved on 02 March 2017 from http://mixedmethodsappraisaltoolpublic.pbworks.com.

Public Health England. (2014). *From evidence into action: opportunities to protect and improve the nation's health.* London: PHE Publications.

Saunders, R., Cape, J., Fearon, P., & Pilling, S. (2016). Predicting treatment outcome in psychological treatment services by identifying latent profiles of patients. *Journal of Affective Disorders, 197*, 107-115.

Appendix D : MMAT flowchart and checklist


Is there a clear research question or objective?
N
Discard
Y
MMAT
Data addresses research question /objective?
N
Discard
Y
RCT
Y
Describes randomisation?
Y N NS □ □ □ 2.1
Describes concealment?
Y N NS □ □ □ 2.2
≥80% complete outcome data?
Y N NS □ □ □ 2.3
<20% drop-out?
Y N NS □ □ □ 2.4
Non-controlled
Y
Recruitment minimises selection bias?
Y N NS □ □ □ 3.1
Measurement appropriate for intervention/outcome?
Y N NS □ □ □ 3.2
Are participants in groups comparable/controlled?
Y N NS □ □ □ 3.3
≥80% complete outcome data and ≥60% response rate (if applic.)?
Y N NS □ □ □ 3.4
Descriptive
Y
Sampling appropriate to quant. research question?
Y N NS □ □ □ 4.1
Representative sample?
Y N NS □ □ □ 4.2
Measurements appropriate?
Y N NS □ □ □ 4.3
≥60% response rate?
Y N NS □ □ □ 4.4
Mixed or qualitative
Y
See next page


Qualitative
Mixed
Is Mixed relevant to both qual. and quant. aspects?
Y N NS
□ □ □
5.1
Is integration of qual. and quant. data/analysis relevant to question/objective?
Y N NS
□ □ □
5.2
Appropriately considered limitations of integration?
Y N NS
□ □ □
5.3
Qual. aspects
Sources of data appropriate to address research question?
Y N NS
□ □ □
1.1
Analysis addresses research question?
Y N NS
□ □ □
1.2
Considered how context affects findings?
Y N NS
□ □ □
1.3
Considered how researcher affects findings?
Y N NS
□ □ □
1.4
Quant. aspects
Non-controlled
RCT
Descriptive
See relevant sections on previous page.

Total score for Mixed is equal to the lowest scoring section. E.g. if Mixed section is 3 (100%), Qual. section is 2 (50%) and Quant. section is 3 (75%), total score is 2 (25%)

Appendix E : Table of theories, their themes, and associated studies

Evidence pertaining to each theory of how PCDSSs worked was gathered from the identified studies and thematically analysed. Each theory, its identified themes, and references for the studies used to identify these themes are given in the table below.

Table E1

Assumed theories of why PCDSSs are effective [pre-analysis], identified themes relating to each theory, and studies themes appear in

#	Theory	Themes (numerical reference)	Study references
1	The problem the PCDSS is being used for can be solved, (the 'right decision' exists)	(1.1) Right for whom?	(Colombet et al., 2003; Sanders et al., 2011)
		(1.2) Point of view	(Colombet et al., 2003; Sanders et al., 2011)
		(1.3) Right evidence	(Chase, 2014; Chorpita et al., 2007; Hunter et al., 2016; Nagpaul, 2001; Wilkinson & Himstedt, 2008)
		(1.4) Right priority	(Colombet et al., 2003; Hunter et al., 2016; Sanders et al., 2011)
		(1.5) Right values	(Hunter et al., 2016; Nagpaul, 2001)
		(1.6) Right place, right action	(Nagpaul, 2001)
		(1.7) Righter/least wrong	(Kennedy et al., 2004; Nagpaul, 2001)
		(1.8) Right is not enough	(Benbenishty & Treistman, 1998; Jenssen et al., 2016)
2	The PCDSS possess enough data to know the right decision for a given individual (both client information and the evidence base are sufficient)	(2.1) Clinicians are necessary	(Chase, 2014; Hunter et al., 2016; Nagpaul, 2001)
		(2.2) The 'Enough Data Paradox'	(Barnett et al., 2002; Cooley et al., 2015; Foster et al., 2014)
		(2.3) Complexity and unusual cases	(Benbenishty & Treistman, 1998)
		(2.4) Context	(Hunter et al., 2016)

		(2.5) Data definitions	(Buckingham et al., 2015; Hunter et al., 2016; Kennedy et al., 2004; Nagpaul, 2001)
		(2.6) Not enough, but better	(Bowles et al., 2014; Carroll et al., 2013a; Foster et al., 2014; Huijbregts et al., 2013; Stallvik et al., 2015; Weisz et al., 2012)
		(2.7) Limits	(Barnett et al., 2002; Buckingham et al., 2015; Chorpita et al., 2007)
3	The PCDSS will produce the right decision (both the model calculations and the technological capabilities are sufficient for the process)	(3.1) Quite complex	(Benbenishty & Treistman, 1998; Chase, 2014; Chorpita et al., 2007; Cooley et al., 2013; Wilkinson & Himstedt, 2008)
		(3.2) Not complex enough	(Benbenishty & Treistman, 1998; Nagpaul, 2001)
		(3.3) Righter	(Benbenishty & Treistman, 1998; Bowles et al., 2014; Foster et al., 2014; Huijbregts et al., 2013; Stallvik et al., 2015; Tolin, Diefenbach, Maltby, & Hannan, 2005; Weisz et al., 2012)
		(3.4) Right is expensive	(Cooley et al., 2013)
		(3.5) Right is simple	(Carroll et al., 2013b; Hunter et al., 2016; Jenssen et al., 2016; Rindal et al., 2013; Sharifi et al., 2014)
		(3.6) Service users are not simple	(Jenssen et al., 2016)
4	The PCDSS produces the right decision more often than a human	(4.1) Better than human	(Benbenishty & Treistman, 1998; Bowles et al., 2014; Carroll et al., 2013a; Foster et al., 2014; Huijbregts et al., 2013; Jenssen et al., 2016; Kennedy et al., 2004; Rindal et al., 2013; Sharifi et al., 2014; Stallvik et al., 2015; Tolin et al., 2011; Weisz et al., 2012)
		(4.2) Doing less with more	(Rollman et al., 2002)
		(4.3) Human trumps tool	(Barnett et al., 2002; Benbenishty & Treistman, 1998; Hunter et al., 2016; Olfson et al., 2003; Sanders et al., 2011)
		(4.4) Low impact on low risk	(Foster et al., 2014; Jenssen et al., 2016)
		(4.5) Idiosyncrasies	(Barnett et al., 2002; Chorpita et al., 2007; Hunter et al., 2016; Nagpaul, 2001; Sanders et al., 2011)

5	Clinicians will want to use the PCDSS	(5.1) What they want can change	(Kennedy et al., 2004; Sanders et al., 2011)
		(5.2) If it fits with work (values and procedures)	(Chase, 2014; Hunter et al., 2016; Sanders et al., 2011)
		(5.3) ...and no further (i)	(Hunter et al., 2016)
		(5.4) If it is a priority	(Olfson et al., 2003; Sanders et al., 2011)
		(5.5) If it is thought to be useful	(Benbenishty & Treistman, 1998; Hunter et al., 2016; Olfson et al., 2003; Sharifi et al., 2014)
		(5.6) If it is believed to be complex enough	(Barnett et al., 2002; Benbenishty & Treistman, 1998; Cooley et al., 2015; Sanders et al., 2011)
		(5.7) ...and no further (ii)	(Lobach et al., 2016)
		(5.8) If it fits clinicians’ definitions of relevant evidence	(Barnett et al., 2002; Olfson et al., 2003; Sanders et al., 2011)
		(5.9) Clinician intuition trumps formal assessment	(Benbenishty & Treistman, 1998; Clarke et al., 2010; Colombet et al., 2003; Wilkinson & Himstedt, 2008)
		(5.10) Do not threaten the clinician	(Benbenishty & Treistman, 1998; Colombet et al., 2003; Kennedy et al., 2004)
		(5.11) Stakeholder buy-in	(Benbenishty & Treistman, 1998; Chase, 2014; Hunter et al., 2016)
		(5.12) Engage with design	(Benbenishty & Treistman, 1998)
		(5.13) Flexible, not rigid	(Barnett et al., 2002; Chorpita et al., 2007; Hunter et al., 2016; van Vliet et al., 2015)
		(5.14) Augment existing procedures	(Chase, 2014)

6	Clinicians will be able to use the algorithm (PCDSS technology, clinician knowledge, and organisational support are sufficient)	(6.1) Integrate with existing (Electronic Health Record) systems	(Huijbregts et al., 2013)
		(6.2) Be user friendly	(Benbenishty & Treistman, 1998; Chase, 2014; Colombet et al., 2003; Jenssen et al., 2016; Lobach et al., 2016)
		(6.3) Clinicians are already user friendly	(Clarke et al., 2010; Jenssen et al., 2016)
		(6.4) Design out ergonomic issues	(Colombet et al., 2003; Kennedy et al., 2004)
		(6.5) Design in organisational support	(Buckingham et al., 2015; Hunter et al., 2016; Kennedy et al., 2004)
		(6.6) Other service pressures	(Benbenishty & Treistman, 1998; Hunter et al., 2016)
		(6.7) Packages of support	(Foster et al., 2014; Kennedy et al., 2004)
		(6.8) Training	(Colombet et al., 2003; Nagpaul, 2001; Sanders et al., 2011)
		(6.9) Phrasing	(Clarke et al., 2010; Colombet et al., 2003)
		(6.10) Routines: friend and foe	(Colombet et al., 2003; Olfson et al., 2003; Sanders et al., 2011)
7	Clinicians will comply with the PCDSS' recommendations	(7.1) High non-adherence to the letter	(Barnett et al., 2002; Benbenishty & Treistman, 1998; Sanders et al., 2011; Wilkinson & Himstedt, 2008)
		(7.2) Higher partial adherence	(Clarke et al., 2010; Huijbregts et al., 2013)
		(7.3) Am I supported to?	(Benbenishty & Treistman, 1998; Chase, 2014; Hunter et al., 2016)
		(7.4) Clinicians will do what they think is best	(Carroll et al., 2013a; Colombet et al., 2003; Nagpaul, 2008; Sanders et al., 2011; Wilkinson & Himstedt, 2008)
		(7.5) If they want to	(Benbenishty & Treistman, 1998; Sanders et al., 2011; Wilkinson & Himstedt, 2008)
		(7.6) If they trust it	(Benbenishty & Treistman, 1998; Colombet et al., 2003; Hunter et al., 2016; Jenssen et al., 2016; Kennedy et al., 2004; Nagpaul, 2001; Olfson et al., 2003; Rollman et al., 2002; Sanders et al., 2011; Sharifi et al.,

			2014)
		(7.7) Treatment adherence improves over baseline	(Clarke et al., 2010; Foster et al., 2014; Jenssen et al., 2016; Sharifi et al., 2014)
		(7.8) Covert improvement	(Clarke et al., 2010)
		(7.9) When adherence high: to assessment	(Carroll et al., 2013a; Clarke et al., 2010; Cooley et al., 2015; Jenssen et al., 2016; Kennedy et al., 2004; Rindal et al., 2013)
		(7.10) When adherence low: to treatment	(Cooley et al., 2015; Foster et al., 2014; Rindal et al., 2013; Sharifi et al., 2014)
8	Clinicians will use the PCDSS in practice	(8.1) First, is it stakeholder designed?	(Benbenishty & Treistman, 1998; Chase, 2014; Colombet et al., 2003; Hunter et al., 2016; Jenssen et al., 2016; Kennedy et al., 2004; Lobach et al., 2016; Rindal et al., 2013; van Vliet et al., 2015)
		(8.2) Does it interfere with what I do already?	(Barnett et al., 2002; Benbenishty & Treistman, 1998; Olfson et al., 2003; Sanders et al., 2011; van Vliet et al., 2015)
		(8.3) Is it supported by seniors?	(Benbenishty & Treistman, 1998; Chase, 2014)
		(8.4) How easily can I ignore it?	(Sharifi et al., 2014)
		(8.5) Will I remember?	(Benbenishty & Treistman, 1998; Olfson et al., 2003; Sanders et al., 2011)
		(8.6) I’m better than a machine	(Benbenishty & Treistman, 1998)
		(8.7) I’ll use it as I like	(Carroll et al., 2013a; Rollman et al., 2002; Wilkinson & Himstedt, 2008)
		(8.8) Competing for attention	(Rollman et al., 2002)

		(8.9) Useful does not mean used	(Colombet et al., 2003; Hunter et al., 2016; Olfson et al., 2003; Sanders et al., 2011; Sharifi et al., 2014)
9	Service users will accept the use of the PCDSS in their care	(9.1) Am I looking for help?	(Huijbregts et al., 2013; Jenssen et al., 2016; Stallvik et al., 2015)
		(9.2) Personal preferences versus guideline recommendations	(Nagpaul, 2001; Sanders et al., 2011)
		(9.3) Poor design needs more help	(Buckingham et al., 2015)
		(9.4) Give me a reason	(Buckingham et al., 2015)
		(9.5) Cultural considerations	(Nagpaul, 2001)
		(9.6) Right context	(Jenssen et al., 2016)
10	Using the PCDSS results in an overall advantage versus standard care (this may be to the organisation, client, clinician, etc.)	(10.1) Treat problems, not diagnoses	(Carroll et al., 2013a; Weisz et al., 2012)
		(10.2) More appropriate care	(Bowles et al., 2014; Carroll et al., 2013a; Foster et al., 2014; Huijbregts et al., 2013; Stallvik et al., 2015; Weisz et al., 2012)
		(10.3) Lower impact on the unusual, co-morbid, and low risk	(Barnett et al., 2002; Benbenishty & Treistman, 1998; Bowles et al., 2014; Foster et al., 2014; Jenssen et al., 2016)
		(10.4) Faster treatment	(Carroll et al., 2013a; Thomas et al., 2004)
		(10.5) Saves more over TAU	(Bowles et al., 2014; Foster et al., 2014; Tolin et al., 2011)
		(10.6) Costs more over nothing	(Sanders et al., 2011)
		(10.7) Highlighting need	(Chase, 2014)
		(10.8) Inter-organisational collaboration	(Chase, 2014)
		(10.9) Inter-organisational conflict	(Sanders et al., 2011)
		(10.10) Time burden	(Colombet et al., 2003; Olfson et al., 2003; Sanders et al., 2011)
		(10.11) Feedback helps me learn	(Benbenishty & Treistman, 1998; Clarke et al., 2010; Hunter et al., 2016; Olfson et al., 2003; Sharifi et al., 2014)

		(10.12) Potentially inappropriate use	(Carroll et al., 2013a; Rollman et al., 2002; Wilkinson & Himstedt, 2008)
		(10.13) No clinician change, no advantage	(Olfson et al., 2003; Rollman et al., 2002)
11	The risks of use of the PCDSS are not significant enough to suggest against its use, and any errors are suitably controlled	(11.1) Errors? What errors?	None (see discussion)
		(11.2) We don't know who is responsible	None (see discussion)
		(11.3) Significance depends on problem	None (see discussion)
		(11.4) Not feasible yet	(Benbenishty & Treistman, 1998; Carroll et al., 2013a; Clarke et al., 2010; Cooley et al., 2015; Olfson et al., 2003; Rindal et al., 2013; Sharifi et al., 2014)
		(11.5) Less risky than clinicians	(Bowles et al., 2014; Carroll et al., 2013b; Foster et al., 2014; Huijbregts et al., 2013; Stallvik et al., 2015; Weisz et al., 2012)
		(11.6) No change, no risk?	(Carroll et al., 2013a; Hunter et al., 2016; Olfson et al., 2003; Rollman et al., 2002; Sharifi et al., 2014)
		(11.7) Less validity for complexity	(Barnett et al., 2002; Benbenishty & Treistman, 1998; Rindal et al., 2013; Sanders et al., 2011; van Vliet et al., 2015)
		(11.8) Limited risk factors	(Benbenishty & Treistman, 1998; Colombet et al., 2003)
		(11.9) Not all with depression are lost (but time and referral ratios are)	(Carroll et al., 2013a; Jenssen et al., 2016)
		(11.10) Culturally insensitive?	(Chorpita et al., 2007; Hunter et al., 2016; Kennedy et al., 2004; Nagpaul, 2001; Sanders et al., 2011)

		(11.11) Good tools are not good clinical judgements	(Carroll et al., 2013a; Nagpaul, 2008)
		(11.12) Interference with ignorability reduce compliance	(Clarke et al., 2010; Jenssen et al., 2016)
		(11.13) Putting clinicians out of a job?	(Wilkinson & Himstedt, 2008)
		(11.14) Competing interests and limited resources	(Colombet et al., 2003; Hunter et al., 2016; Olfson et al., 2003; Sanders et al., 2011)
		(11.15) Trees will die	(Bowles et al., 2014; Buckingham et al., 2015; Carroll et al., 2013a; Tolin et al., 2011)
12	The PCDSS produces better outcomes because its decisions are more researched-based and less biased than a clinicians		This theory was tested using a summary of all the research, which is expanded on in Appendix G.12

Appendix F : Table of hypotheses, their chains of inference, and associated themes

Hypotheses on conditions of PCDSS effectiveness, and their derivative chains and themes (post-analysis). All identified studies can be linked to an associated hypothesis via themes (see previous Table E1)

Table F1

Table of hypotheses, their chains of inference, and associated themes

Hypotheses (mechanism, context, and outcome chains)	**Chain of inference**	**References to themes from the literature**
[1] Involvement of stakeholders early in the PCDSS design process is important to: i. increase organisational fit (values, priorities, processes) ii. integrate appropriate evidence sources (local, clinician, research, service users) iii. reduce ergonomic flaws iv. increase stakeholder ownership all of which improve the chance of PCDSS use	Managerial support is need to prioritise resources for a PCDSS	(1.1) (5.4) (5.11) (7.4) (8.3) (8.8)
	Value, context, and evidence can all be designed in	(2.5) (5.2) (5.8) (5.12) (5.13) (5.14) (6.1) (6.5) (8.1) (8.2)
	Tools that are easy to use (at a technological and service level) are more likely to be used	(5.14) (6.1) (6.4) (6.7 – 6.10) (8.2) (9.3)
	Use is partly dependent on how easy the tool is to be use, which can be designed in	(6.1) (6.2) (6.4) (6.9) (8.4) (8.5) (8.9) (9.3) (10.10) (11.12)
[2] PCDSSs improve outcomes: i. for services when they use limited resources more efficiently, and improve inter-agency working ii. for clients by selecting more effective treatment for problems, increasing recovery rate, and getting	Improves services as well as client outcomes	(10.7) (10.8)
	Unavoidable costs of time and resources	(10.10) (11.15)
	Success is partly resource-driven: are enough available? Does their use decrease overall?	(3.4) (6.6) (7.4) (8.8) (10.10) (10.5) (10.6) (10.9)

treatment faster, in conjunction with a higher diagnostic rate iii. for clinicians by making decisional models explicit, which increases chances of adhering to guidelines	When assessment with a PCDSS occurs, more likely to identify cases for intervention earlier	(7.9) (10.4)
	When assessment with the PCDSS occurs, there is an increased chance appropriate treatment being offered	(7.2) (7.5) (7.7) (10.2)
	Improved identification of clinical symptoms alone increases costs by increasing the number of unwanted referrals for clients	(9.1) (9.2) (10.1) (10.6) (11.9)
	More appropriate treatment saves money by decreasing time in treatment through increased effectiveness	(4.2) (10.1) (10.2) (10.5)
[3] Impact on mental health outcomes depends on (in order of importance): i. change in clinician behaviour ii. support from the organisation in which the PCDSS is used iii. a broad and well-integrated evidence base	Models are averages, so human judgement is necessary for wider success of decisions	(2.2) (2.4) (4.4) (4.5)
	Outcomes are more likely to improve when clinicians work with a PCDSS (rather than the PCDSS working on the clinician)	(1.8) (2.1) (2.2) (2.3) (2.6) (2.7) (3.1) (3.2) (3.5) (3.6) (4.1) (4.5) (5.10) (10.11)
	When assessment with the PCDSS occurs, there is an increased chance appropriate treatment being offered	(7.2) (7.5) (7.7) (10.2)
	Clinicians are in a position of power regarding the success of a PCDSS	(4.3) (5.2) (5.3) (5.5 – 5.8) (5.10) (6.3) (7.4) (7.5) (8.6) (8.7) (10.12) (10.13)
	Managerial support is need to prioritise resources for a PCDSS	(1.1) (5.4) (5.11) (7.4) (8.3) (8.8)
	Success is partly resource-driven: are enough available? Does their use decrease overall?	(3.4) (6.6) (7.4) (8.8) (10.10) (10.5) (10.6) (10.9)

	Value, context, and evidence can all be designed in	(2.5) (5.2) (5.8) (5.12) (5.13) (5.14) (6.1) (6.5) (8.1) (8.2)
	Tools that are easy to use (at a technological and service level) are more likely to be used	(5.14) (6.1) (6.4) (6.7 – 6.10) (8.2) (9.3)
	When assessment with a PCDSS occurs, more likely to identify cases for intervention earlier	(7.9) (10.4)
[4] PCDSSs are more likely to be used and adhered to when they are trusted as decision makers at a level approaching that of an individual clinician with regard to their ability to account for: i. client complexity (including co-morbidity and additional risk factors) ii. organisational context (including values, priorities and procedures)	Use is not related to increased compliance with recommendations	(7.1) (7.4) (7.5) (7.10) (8.6) (8.7)
	Flexibility of implementation improves use but decreases absolute adherence to recommendations	(5.9) (7.1) (7.2) (7.4) (7.7) (7.10) (8.6)
	Clinicians are in a position of power regarding the success of a PCDSS	(4.3) (5.2) (5.3) (5.5 – 5.8) (5.10) (6.3) (7.4) (7.5) (8.6) (8.7) (10.12) (10.13)
	Clinicians have inherently low trust in the tool compared to themselves	(3.2) (4.3) (5.9) (7.1) (7.6) (8.7)
[5] It is more important for a PCDSS to make valued decisions according to the context it is employed in than decisions that are right according to the research base, in order to produce the most effective mental health outcomes	Unavoidable costs of time and resources	(10.10) (11.15)
	An increase in valued outcomes is a more acceptable focus than the ‘right’ outcome	(1.5) (1.8) (2.4) (3.6) (7.5)
	Value depends on context	(1.1) (1.5) (2.4) (5.2) (5.3)
	The ‘right’ solution is always going to be a matter of dispute	(1.1 – 1.8) (9.2) (9.5)

	Success is partly resource-driven: are enough available? Does their use decrease overall?	(3.4) (6.6) (7.4) (8.8) (10.10) (10.5) (10.6) (10.9)
[6] Trust in PCDSSs is related to their ability to: i. decrease perception of risk to the clinician and organisation ii. integrate appropriate sources of information, including local-, research-, and client-based iii. do so transparently, including degree of uncertainty and risk factors for treatment iv. allow clinicians to exercise their own decisional discretion v. personalise decisions to client and context	Unavoidable costs of time and resources	(10.10) (11.15)
	Responsibility for PCDSS decisions needs to be clearly defined in order to manage risk	(8.7) (10.12) (10.13) (11.1) (11.2) (11.3) (11.6)
	Simple decisions increase absolute adherence; complex decisions (with more options) decrease absolute adherence	(4.3) (7.1) (7.2) (7.6) (7.9) (7.10)
	Trust does not rely on the research base	(2.3) (4.3) (5.9) (5.10) (7.6)
[7] PCDSSs are most effective at improving mental health outcomes when they function as part of a mutually symbiotic relationship between clinician, organisation, and tool in order to integrate expert, contextual, and research knowledge to a high degree of complexity	The 'right' solution is always going to be a matter of dispute	(1.1 – 1.8) (9.2) (9.5)
	Clinicians can mediate the risk of inappropriate decisions made by a PCDSS	(2.7) (5.13) (9.3) (9.4) (10.3) (10.9) (11.4) (11.7 – 11.11) (11.13) (11.14)
	Models are averages, so human judgement is necessary for wider success of decisions	(2.2) (2.4) (4.4) (4.5)
	Outcomes are more likely to improve when clinicians work with a PCDSS (rather than the PCDSS working on the clinician)	(1.8) (2.1) (2.2) (2.3) (2.6) (2.7) (3.1) (3.2) (3.5) (3.6) (4.1) (4.5) (5.10) (10.11)
	Support for the PCDSS needs to be encouraged on multiple fronts	(4.3) (5.1) (5.2) (5.5) (5.8) (5.10) (5.11) (5.14) (7.3) (7.4)
	A PCDSS can decrease the risk of clinicians making inappropriate decisions	(2.6) (3.3) (4.1) (7.8) (10.2) (11.5)

[8] PCDSSs are more likely to improve mental health outcomes when they are matched to specific contexts and problems, rather than deployed generically	Matching context to referral/treatment recommendation can increase perception of PCDSS usefulness in the client	(9.1) (9.6)
	An increase in valued outcomes is a more acceptable focus than the ‘right’ outcome	(1.5) (1.8) (2.4) (3.6) (7.5)
	Value depends on context	(1.1) (1.5) (2.4) (5.2) (5.3)

Appendix G : Narrative summaries of the evidence for each theory

The initial theories on how PCDSSs work were the driver for the rest of realist synthesis, ultimately leading to the hypotheses. The theories have here been updated given the information from the analysis and are included for completeness.

G.1 Theory 1: The problem the PCDSS is being used for can be solved, (the 'right decision' exists)

What is 'right' is a subjective judgement that inherently produces conflict in all but perhaps the simplest of cases. Any decision—whether PCDSS- or clinician-derived—is therefore biased towards a certain point of view, which may be more suited to solving the problems of the clinician, the work organisation, the service user, or indeed the researcher. It is thereby important to ask 'for whom and on what basis is this the right decision?'. For instance, referral to a specialist for the treatment of depression may be right for the client, but not for an over-subscribed service. The intervention on offer may be right according to the research-base, but not according to a particular client's circumstances. The best that can be hoped for is a 'righter' decision, bearing in mind 'usefulness' will likely supersede 'rightness' in practice. Ultimately, choices must be made to balance priorities, evidence sources, and values among other considerations to ensure the best ecological fit between PCDSS and problem.

G.2 Theory 2: The PCDSS possess enough data to know the right decision for a given individual (both client information and the evidence base are sufficient)

On their own, the PCDSS designs possess enough data to make more appropriate decisions, but are unable to do so with absolute accuracy. This is affected by the level of integration with organisational information (relevant priorities, procedures, referral partners, etc.), and how specific the PCDSS is to the client’s problems. A tool that separately addresses common problems in ADHD is more likely to succeed than one looking at ADHD in general, for instance.

PCDSSs are less likely to make correct decisions in situations relying heavily on context (e.g. organisational values, location-specific legal definitions), complexity (e.g. co-morbid diagnoses), or unusual cases, as research evidence is usually lacking and/or low in ecological validity. Paradoxically, the more information the PCDSS possess the less evidence-based its decision, as exponentially less research is available to draw on. Clinicians are therefore necessary to interpret tool decisions: the tool is unlikely to posses enough data on its own, and is likely to be limited in its ability to process such information for the foreseeable future.

G.3 Theory 3: The PCDSS will produce the right decision (both the model calculations and the technological capabilities are sufficient for the process)

The studies demonstrate a level of technological advancement capable of handling sophisticated models produced from a variety of evidence sources. PCDSSs are rarely used for complex data input—they are instead often limited to a single simple questionnaire—so it is unknown whether they are technologically capable of computing large numbers of variables. However, the evidence suggests current PCDSSs are able to produce ‘righter’ decisions from quite complex models that are sufficient for assessment and treatment decisions. More complex models are associated with an exponential increase in development costs. The right decision is easier to produce when choices are straightforward and logic-driven (e.g. ‘if smoker, offer counselling’), but harder when individual variation is a stronger factor (e.g. ‘what treatment will Molly most prefer?’). The professional consensus is that neither model nor technology will ever be enough to produce the right decision for every circumstance. As a model appears to require stakeholder knowledge in order to make appropriately complex decisions, and this knowledge is inherently biased, it appears unlikely a model can be developed that always produces the right decision in all but the simplest circumstances.

G.4 Theory 4: The PCDSS produces the right decision more often than a human

For most decisions, a PCDSS will perform better than a human in respect to making more appropriate therapy decisions for a given problem/diagnosis. This is more likely for high-risk clients, whereas there is less-to-negligible impact on low-risk clients. Better outcomes are not sufficiently explained by a higher dose of therapy, as often the tool prescribes a similar amount of treatment, if not less. Better matching of problem to treatment and more frequent adjustments to therapy are more likely to mediate this effect.

PCDSSs are less able to make the right decision when idiosyncratic factors are prominent, such as organisational values, individual client preferences, or client/clinician/workplace culture. This is due to a smaller evidence base for the tool to draw on, although the ‘knowledge gap’ can be reduced with integration of stakeholder knowledge in the construction of PCDSS models. However, partly due to its reduced capacity to deal with such cases, clinicians will usually value their own decision over the tool, despite evidence of PCDSS efficacy. The ‘right’ decision is therefore mostly regarded as the human one.

G.5 Theory 5: Clinicians will want to use the PCDSS

Whether clinicians want to use a PCDSS depends on their perception of several professional issues: does it fit with the values and contingencies of my workplace? Is the decisional area of the tool a clinical priority? Does it incorporate what I think is relevant evidence and do this with sufficient complexity? Is it useful to me?

The 'default position' for most clinicians is an assumed 'no': the PCDSS is unlikely to fit with their organisational values, it is probably not complex enough, and (therefore) unlikely to be useful. This is supported by the entrenched belief that clinical intuition is superior to formalised assessment processes, so a prescriptive PCDSS is inherently less valuable. This means clinicians are generally not in favour of PCDSS use; a view that is resistant to change despite the presentation of evidence contradicting their position on any single issue. However, practitioners are more likely to want to use a PCDSS if its suitability can be demonstrated on multiple fronts. This includes backing from senior management (values), stakeholder involvement in design processes (complexity), inclusion of valued evidence sources (evidence), and augmentation with existing procedures (values). A flexible tool incorporating the user's expertise—rather than rigidly excluding it—is less likely to be perceived as threatening that position of expertise, and more likely to be accepted.

G.6 Theory 6: Clinicians will be able to use the tool (PCDSS technology, clinician knowledge, and organisational support are sufficient)

PCDSSs are easier to use if they are integrated into existing (and thereby familiar) systems, such as Electronic Health Records. Usability techniques—including use of colour, visuals, and reminders—can improve the ability of clinicians to use a tool quickly and reduce learning time. However, it is difficult to make a tool more ‘user-friendly’ than informal techniques already used by clinicians as humans are, by definition, quite adept and familiar with performing such tasks themselves. This means practitioners are more likely to lapse into informal processes when offered the choice between those and formal tool-based ones without further motivation. Offering a package of ongoing support can grant opportunities to address usability issues, such as educational sessions, regular audits, feedback sessions, technical assistance and frequent contact with the (research) team. PCDSS training is helpful, especially if the tool is used infrequently and training can be offered at multiple time-points.

Designing out ergonomic issues (speed, intuitiveness, etc.) with user testing early in the design process is important to improve usability. Understanding of question items can also be checked with users, as certain items, especially around social and psychological factors, are more open to interpretation.

Support from the organisation(s) in which the tool is to be deployed is also essential, as service pressures alone can prevent tool adoption. Support can be enhanced by designing the PCDSS with key stakeholders from the beginning to improve fit and ownership. This can also help integrating the tool with existing organisational routines, which increases likelihood of use. However, if the new procedures are too similar to existing but ineffective ones, the PCDSS is more likely to be ignored.

G.7 Theory 7: Clinicians will comply with the PCDSS’ recommendations

Compliance is linked to trust in the efficacy of the tool, in particular that its decision-making is more effective than the using clinician’s. Trust in the tool tends to be low by default, while trust in the clinician’s own abilities is high, so few PCDSSs demonstrate good adherence.

Ultimately, practitioners will do what they think is best professionally, therefore adherence to recommendations in totality is often low. However, partial adherence (picking and choosing recommendations to follow) is much higher, where clinicians can exercise their own judgement more. Recommendations that are rigid are more likely to (be seen to) interfere with established procedures of both the practitioner and their workplace, reducing compliance.

The principles of trust, rigidity and partial adherence are demonstrated through compliance with treatment and assessment decisions. Treatment decisions are more complex, so trust in the PCDSS to make them is lower, but there are more opportunities to exercise discretion over which parts of the recommendations to follow. Complete adherence therefore tends to be low, but guideline compliance overall improves compared to baseline, as clinicians are more likely to take the PCDSS's decision into account when making their judgement. Compliance with assessment shows the opposite pattern: to assess or not is a simpler decision, so trust in the PCDSS is higher, but there is no room for partial adherence. When compliance occurs it is normally total. However, the rigidity is more likely to be seen as an interference, which can result in a drop in assessments compared to baseline.

Perceptions of trust and interference can be improved with support from colleagues, senior managers, and organisational procedures. Compliance tends to be lower when these are lacking, and higher when present.

G.8 Theory 8: Clinicians will use the PCDSS in practice

Practitioners tend not to use PCDSSs in practice due to low levels of trust in the tool compared to their own judgement, and/or poor integration with their work context. Use can be enhanced by taking into account several practical measures, starting with stakeholder involvement early in the design process. This enhances the potential for good organisational fit, sense of ownership, and usability. Generally, the more a tool interferes with existing practice the less likely it is to be used, while continuous user-driven improvement enhances the chances of permanent use. The exception to the rule of low interference is when a PCDSS can be easily ignored, as this reduces the chance of behavioural change.

Support by senior management is important for implementation and longevity, especially as other priorities will compete for clinician attention. Reminders to use the tool are helpful to sustain use, especially when a particular PCDSS is used infrequently. Clinicians are more likely to use the PCDSS if its decisions are flexible, responsive to the organisational and client contexts, and the overall design is user-friendly. Practitioners will tend to use the PCDSS in the way they want to, so an overly-prescriptive tool is more likely to be dropped. When gathering feedback, care should be taken not to mistake comments that the tool is 'useful' to mean that it will be used in practice, as this is rarely a sufficient condition.

G.9 Theory 9: Service users will accept the use of the PCDSS in their care

PCDSSs are generally acceptable to clients, although adherence to decisions are more likely when they are actively seeking help. For instance, a referral for depression is more likely to result in a client turning up for treatment if they were originally looking for support for symptoms, than if depression is identified from an otherwise unrelated opportunistic screening. Clients are more likely to consider a recommendation if it is made within a salient context; for instance referrals for smoking cessation are more readily accepted when a paediatrician, rather than a GP, says it improves the health of your child. As with clinicians, acceptance of recommendations is also more likely when personal preferences can be factored in, instead of being ridden roughshod by rigid guidelines.

Poorly designed tools require more input by clinicians in order to make them acceptable. This applies both to understanding how to use a tool (e.g. questionnaire input) and why it is helpful to themselves. Use and justification can be made more obvious with appropriate user testing and stakeholder involvement in design. The later is especially important to integrate cultural considerations into decisions, which can have a significant impact on treatment acceptability.

G.10 Theory 10: Using the PCDSS results in an overall advantage versus standard care (this may be to the organisation, client, clinician, etc.)

PCDSS use alone is associated with faster treatment gains (even in the absence of treatment variation), but the difference compared to controls tends to reduce over time. Increases in the appropriateness of treatment (and not the amount of treatment per se) are more likely to see clients maintain those gains. Resource expenditure therefore tends to be less compared to treatment as usual, although this is not the case if no treatment is normally offered. Comparative savings thus depend on existing procedures for a given organisation. All PCDSSs are likely to require more time to use than current practice, particularly when it replaces informal assessment with formal protocols.

Gains are more likely to be seen when problems rather than diagnoses are treated, as treatment is more targeted. Gains are less likely for low-risk, unusual, co-morbid, or sub-threshold cases, although this can be improved with greater clinician discretion in the decision outcome. However, such flexibility also increases the potential for misuse of tools and any materials associated with them. This can be moderated by making clear the decision-making logic, as this helps the clinician reflect on their own judgements, which can enhance adherence to guideline recommendations. Ultimately, no advantage will be seen if the practitioner does not alter their behaviour.

One of the biggest advantages of PCDSSs over standard care is the opportunity for them to highlight need within the service, particularly for further decisional support, and promote inter-agency collaboration. A tool regularly informed by such multi-disciplinary thinking can make better decisions and improve cross-organisational procedures. On the flip side, a poorly informed tool can increase inter-organisational conflict through poor (within an organisational context, not necessarily a research one) referral decisions.

G.11 Theory 11: The risks of use of the PCDSS are not significant enough to suggest against its use, and any errors are suitability controlled

There is little direct evidence or acknowledgement of PCDSS errors, making it difficult to draw conclusions for this theory. Instead, more questions are raised that deserve consideration:

Who is responsible for PCDSS mistakes? Practitioners potentially assume the responsibility is theirs, which can then reduce PCDSS use, as risks to themselves are seemingly greater (particularly if decisional mechanisms are unclear).

What is an acceptable level of error? Is it equal to or less than a human’s? If it is equal, does further risk management need to happen, given it does not for a clinician? The significance of errors should also depend on the problem under consideration, e.g.

decisions on potential for suicide are inherently more risky if they are wrong than those on screening for anxiety.

There are many risks to using PCDSSs, although these can be modified with appropriate consideration. Most prevalent is that the current level of evidence and technology means PCDSSs are less valid for certain cases, particularly complex ones, so it is not feasible to deploy them without a clinician to give input on decisions. Evidence on risk for interventions is also low, which again requires a clinician to monitor and adjust for inappropriate judgements. This means no PCDSS should be putting practitioners out of a job (just yet), and using a tool without expert knowledge will increase the risk of poor decisions. Over-reliance on tools is a risk in modern practice, particularly when they are potentially cheaper than their human counterparts.

PCDSSs help identify clinical cases otherwise missed by practitioners, yet also increase the number of false-positives or those inappropriate for referral. This can impact diagnosis:referral ratios and waste clinical time, as well as pressuring service users to take unwanted treatment. Clinician discretion is again helpful, as well as appropriate integration of patient preferences into PCDSS models.

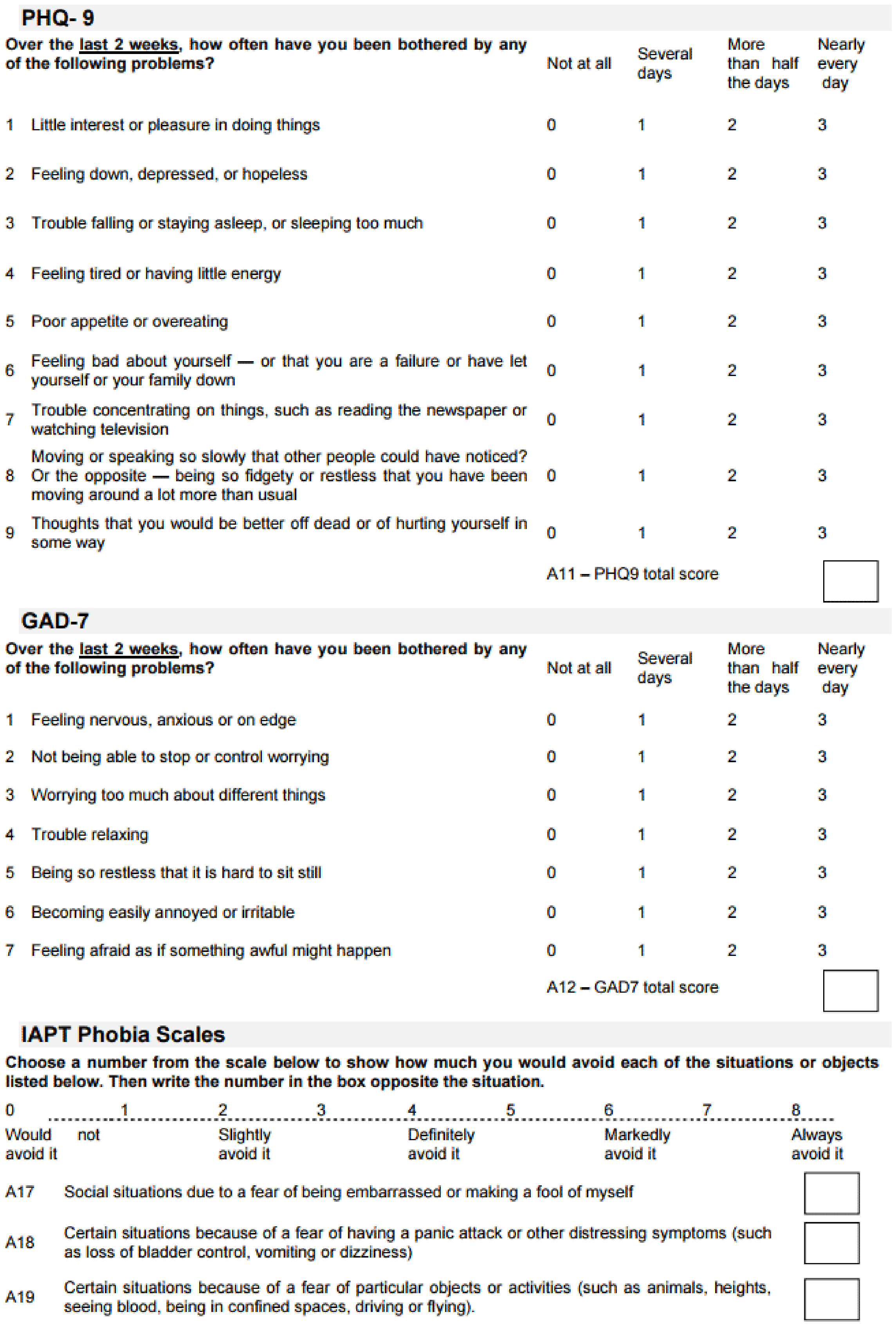
Usually PCDSSs are less risky than clinicians, largely due to their higher prescription of appropriate treatment. However, when tools are seen as interfering with work and are easily ignored, they can worsen practice. PCDSSs also compete for resources with other treatments, which need to be accounted for in their design to reduce the risk of inappropriate referrals, resource exhaustion, etc. This extends to use of natural resources as well, since many tools involve a paper-based element, such as questionnaires. Good integration with electronic technologies can reduce the impact on the environment.

G.12 Theory 12: The PCDSS produces better outcomes because its decisions are more research-based and less biased than a clinician’s

A PCDSS depends upon being used by a clinician and supported by the host organisation in order to effect change. These are necessary, although not sufficient, conditions for better outcomes. Use and support can be enhanced through appropriate stakeholder design and usability testing.

PCDSSs produce better outcomes partly because they address biases, rather than avoid them. Incorporating a wealth of local and expert evidence such as organisational priorities and referral criteria—which are not necessarily experimentally informed—in addition to the research base creates a more valued PCDSS that suits its context. This also helps address clinician beliefs that they are better decision-makers by demonstrating appropriate complexity and contextual awareness. Nevertheless, a PCDSS will produce the best outcomes if it works with clinicians, and has the flexibility for them to exercise their own expertise and discretion. This allows a) better management of unusual, co-morbid, or below-threshold cases, and b) reflection on clinician judgement processes. The latter can be enhanced through transparent PCDSS reasoning models, as they help organically bring practitioner decisions more in-line with guideline recommendations.

Appendix H : IAPT Minimum Data Set



PHQ- 9

Over the last 2 weeks, how often have you been bothered by any of the following problems?		Not at all	Several days	More than half the days	Nearly every day
1	Little interest or pleasure in doing things	0	1	2	3
2	Feeling down, depressed, or hopeless	0	1	2	3
3	Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4	Feeling tired or having little energy	0	1	2	3
5	Poor appetite or overeating	0	1	2	3
6	Feeling bad about yourself — or that you are a failure or have let yourself or your family down	0	1	2	3
7	Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
8	Moving or speaking so slowly that other people could have noticed? Or the opposite — being so fidgety or restless that you have been moving around a lot more than usual	0	1	2	3
9	Thoughts that you would be better off dead or of hurting yourself in some way	0	1	2	3

A11 – PHQ9 total score □

GAD-7

Over the last 2 weeks, how often have you been bothered by any of the following problems?		Not at all	Several days	More than half the days	Nearly every day
1	Feeling nervous, anxious or on edge	0	1	2	3
2	Not being able to stop or control worrying	0	1	2	3
3	Worrying too much about different things	0	1	2	3
4	Trouble relaxing	0	1	2	3
5	Being so restless that it is hard to sit still	0	1	2	3
6	Becoming easily annoyed or irritable	0	1	2	3
7	Feeling afraid as if something awful might happen	0	1	2	3

A12 – GAD7 total score □

IAPT Phobia Scales

Choose a number from the scale below to show how much you would avoid each of the situations or objects listed below. Then write the number in the box opposite the situation.

0	1	2	3	4	5	6	7	8
Would not avoid it		Slightly avoid it		Definitely avoid it		Markedly avoid it		Always avoid it

A17 Social situations due to a fear of being embarrassed or making a fool of myself □

A18 Certain situations because of a fear of having a panic attack or other distressing symptoms (such as loss of bladder control, vomiting or dizziness) □

A19 Certain situations because of a fear of particular objects or activities (such as animals, heights, seeing blood, being in confined spaces, driving or flying). □

IAPT Employment Status Questions

A14 - Please indicate which of the following options best describes your current status:

Employed full-time (30 hours or more per week)	□
Employed part-time	□
Unemployed	□
Full-time student	□
Retired	□
Full-time homemaker or carer	□

A15 - Are you currently receiving Statutory Sick Pay?

Yes	□
No	□

A16 - Are you currently receiving Job Seekers Allowance, Income support or Incapacity benefit?

Yes	□
No	□

Work and Social Adjustment

People's problems sometimes affect their ability to do certain day-to-day tasks in their lives. To rate your problems look at each section and determine on the scale provided how much your problem impairs your ability to carry out the activity.

1. **WORK** - if you are retired or choose not to have a job for reasons unrelated to your problem, please tick N/A (not applicable)

0	1	2	3	4	5	6	7	8	N/A
Not at all		Slightly		Definitely		Markedly	Very severely, I cannot work		□

2. **HOME MANAGEMENT** – Cleaning, tidying, shopping, cooking, looking after home/children, paying bills etc

0	1	2	3	4	5	6	7	8
Not at all		Slightly		Definitely		Markedly	Very severely	

3. **SOCIAL LEISURE ACTIVITIES** - With other people, e.g. parties, pubs, outings, entertaining etc.

0	1	2	3	4	5	6	7	8
Not at all		Slightly		Definitely		Markedly	Very severely	

4. **PRIVATE LEISURE ACTIVITIES** – Done alone, e.g. reading, gardening, sewing, hobbies, walking etc.

0	1	2	3	4	5	6	7	8
Not at all		Slightly		Definitely		Markedly	Very severely	

5. **FAMILY AND RELATIONSHIPS** – Form and maintain close relationships with others including the people that I live with

0	1	2	3	4	5	6	7	8
Not at all		Slightly		Definitely		Markedly	Very severely	

A13 – W&SAS total score □

Appendix I : Definitions of reliable recovery, improvement, deterioration and change, based on IAPT (2014)

Symptom scores on the PHQ9 and GAD7 usually vary between two time points. This change may reflect random variation due natural error in the measurements or a 'true' change in level of clinical symptoms. This 'true' change is referred to as 'reliable change', which may be a decrease in scores (reliable improvement) or an increase (reliable deterioration). Each measurement has its own Reliable Change Index:

Reliable Change Index

PHQ9: score change ≥6

GAD7: score change ≥4

Clinical cut-off points can also be taken into consideration. When someone enters IAPT with scores indicating clinically significant levels of depression and/or anxiety (i.e. 'at caseness'), and then leaves IAPT with scores below the cut-off after finishing a course of treatment, they may be said to have recovered. If this change is large enough to be classified as reliable improvement, they may be considered to have reliably recovered:

Caseness Threshold

PHQ9: score ≥10 (indicates a diagnosis of depression)

GAD7: score ≥8 (indicates a diagnosis of anxiety)

Both PHQ9 and GAD7 must be below caseness with at least one reliable change to count as recovery.

In IAPT, whether change is reliable, in what direction, and whether this constitutes recovery depends on a combined analysis of PHQ9 and GAD7 scores, shown in the Table below.

Table I1

Guide to calculating clinical outcomes based on PHQ9 and GAD7 scores

Indication	**Measure**	
	PHQ9	GAD7
Reliable improvement	✓	✓
	✓	–
	–	✓
No change	✓	✗
	✗	✓
	–	–
Reliable deterioration	✗	✗
	–	✗
	✗	–

✓ denotes reliable reduction in measure score

✗ denotes reliable increase in measure score

– denotes no reliable change in measure score

Appendix J : Characteristics and outcomes for ML and MAP profiles

Table J1

Summary of ML profile characteristics and outcomes on the MDS based on cases in the 2009-2013 dataset; standard deviations are given in brackets

	Full sample	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Proportion of sample	100%	18.0%	19.1%	3.7%	4.1%	10.2%	9.7%	13.1%	22.2%
Age: mean	38.4 (13.6)	33.3 (8.2)	30.3 (7.2)	65.8 (10.0)	64.9 (9.3)	53.8 (8.0)	40.1 (9.2)	42.9 (9.6)	29.6 (6.7)
Gender: proportion female	66%	64%	70%	67%	69%	64%	57%	59%	72%
Ethnicity: proportion non-white	26%	22%	21%	15%	14%	22%	29%	30%	32%
Welfare: proportion on benefits	30%	12%	8%	9%	11%	44%	54%	73%	26%
Medication: proportion prescribed	42%	27%	21%	32%	37%	57%	62%	75%	41%
Phobia: proportion with phobia	44%	19%	31%	18%	29%	48%	46%	87%	53%
Depression: PHQ9 mean*	14.2 (6.7)	5.6 (3.1)	11.4 (3.1)	5.0 (3.2)	11.1 (3.2)	18.2 (3.5)	13.5 (3.4)	23.0 (2.8)	18.9 (3.2)
Anxiety: GAD7 mean*	12.6 (5.5)	5.6 (2.8)	13.0 (2.8)	4.2 (2.8)	11.2 (3.2)	16.0 (2.9)	8.2 (2.7)	18.5 (2.5)	16.5 (2.8)
Functioning: WSAS mean*	18.6 (9.9)	9.3 (6.1)	14.7 (6.3)	7.6 (6.6)	12.8 (7.2)	18.8 (7.6)	21.3 (6.8)	32.0 (5.7)	22.6 (2.8)
Reliable recovery**	33%	43%	43%	45%	45%	30%	34%	17%	30%
Reliable improvement**	59%	58%	64%	54%	62%	60%	48%	50%	63%
Reliable deterioration**	8%	9%	8%	13%	8%	7%	19%	5%	6%
Dropout**	40%	33%	38%	22%	26%	38%	39%	43%	47%

*Higher scores indicate greater impairment. The following cut-offs may be used: PHQ9—score of 10-14 (moderate), 14-19 (moderately severe), ≥20 (severe) (Kroenke & Spitzer, 2002); GAD7—score of 8-14 (moderate), ≥15 (severe) (Kroenke, Spitzer, Williams, Monahan, & Löwe, 2007); WSAS—score of 10-20 (moderately impaired), ≥21 (severely impaired) (Mundt et al., 2002)

**Of those entering IAPT at caseness; reliable statistics are based on service users who complete a course of treatment

Table J2

Summary of MAP profile characteristics on the MDS based on all cases in the 2009-2013 dataset; standard deviations are given in brackets

	Full sample	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Proportion of sample	100%	16.9%	21.3%	3.1%	4.2%	9.0%	9.0%	14.2%	22.4%
Age: mean	38.4 (13.6)	33.7 (8.5)	30.7 (7.5)	67.0 (9.7)	65.7 (9.1)	54.5 (7.9)	40.8 (9.2)	43.3 (9.5)	30.0 (7.0)
Gender: proportion female	66%	64%	70%	66%	70%	64%	56%	59%	71%
Ethnicity: proportion non-white	26%	22%	22%	15%	14%	21%	29%	30%	33%
Welfare: proportion on benefits	30%	12%	9%	9%	10%	43%	55%	74%	27%
Medication: proportion prescribed	42%	28%	22%	31%	36%	57%	63%	76%	42%
Phobia: proportion with phobia	44%	18%	30%	18%	29%	47%	46%	87%	45%
Depression: PHQ9 mean	14.2 (6.7)	5.3 (3.0)	11.2 (3.2)	4.7 (3.1)	10.7 (3.6)	18.2 (3.5)	13.4 (3.5)	23.1 (2.8)	18.9 (3.2)
Anxiety: GAD7 mean	12.6 (5.5)	5.3 (2.7)	12.6 (3.0)	3.7 (2.5)	10.8 (3.3)	16.0 (2.9)	8.0 (2.7)	18.5 (2.5)	16.5 (2.8)
Functioning: WSAS mean	18.6 (9.9)	9.0 (6.0)	14.7 (6.3)	7.2 (6.5)	12.3 (7.2)	18.7 (7.6)	21.4 (6.7)	32.1 (5.7)	22.7 (7.1)
Reliable recovery*	33%	43%	43%	47%	45%	31%	34%	17%	29%
Reliable improvement*	59%	58%	63%	54%	60%	61%	48%	50%	63%
Reliable deterioration*	8%	9%	8%	16%	9%	7%	19%	5%	6%
Dropout*	40%	33%	38%	22%	25%	37%	38%	43%	47%

*Of those entering IAPT at caseness; reliable statistics are based on service users who complete a course of treatment

Please note that the above MAP figures will differ slightly from those given in Saunders et al. (2016) due to small differences in calculations (this study used SPSS and LatentGOLD for analyses, whereas Saunders used Mplus) and trivial variations in sampling.

Appendix K : Logistic regression outputs examining reliable recovery for MAP, ML, mBCH, and SP models

Where the odds ratio is greater than one, recovery is more likely with high-intensity therapy. Where the odds ratio is less than one, recovery is more probable with low-intensity. Please note some of the information for significant predictors in the MAP and ML tables are also found in Table 6.

Table K1

Logistic regression output for MAP profiles and therapeutic intensity as predictors of reliable recovery, split by profile

						95% Confidence interval for odds ratio	
Profile	**Variable**	**Beta**	**Standard error**	**Significance**	**Odds ratio**	**Lower bound**	**Upper bound**
1	Intercept	.037	.122	.761			
	Therapy intensity	.227	.148	.125	1.254	.939	1.675
2	Intercept	-.050	.062	.420			
	Therapy intensity	.271	.074	.000*	1.311	1.133	1.516
3	Intercept	.087	.295	.768			
	Therapy intensity	-.221	.420	.600	.802	.352	1.827
4	Intercept	.144	.120	.231			
	Therapy intensity	-.063	.155	.685	.939	.693	1.272
5	Intercept	.623	.089	.000			
	Therapy intensity	-.013	.114	.912	.987	.790	1.234
6	Intercept	.305	.082	.000			
	Therapy intensity	.430	.112	.000*	1.537	1.235	1.914
7	Intercept	1.381	.072	.000			
	Therapy intensity	.236	.120	.048*	1.266	1.002	1.600
8	Intercept	.608	.061	.000			
	Therapy intensity	.203	.077	.009*	1.226	1.053	1.427

*odds ratio is significant to p<0.05

Table K2

Logistic regression output for ML profiles and therapeutic intensity as predictors of reliable recovery, split by profile

						95% Confidence interval for odds ratio	
Profile	**Variable**	**Beta**	**Standard error**	**Significance**	**Odds ratio**	**Lower bound**	**Upper bound**
1	Intercept	-.006	.108	.957			
	Therapy intensity	.267	.131	.043*	1.305	1.009	1.689
2	Intercept	-.037	.066	.575			
	Therapy intensity	.252	.079	.001*	1.287	1.103	1.502
3	Intercept	.150	.245	.542			
	Therapy intensity	-.172	.322	.595	.842	.448	1.585
4	Intercept	.114	.120	.339			
	Therapy intensity	-.046	.155	.767	.955	.704	1.295
5	Intercept	.604	.087	.000			
	Therapy intensity	.060	.111	.591	1.062	.854	1.320
6	Intercept	.297	.080	.000			
	Therapy intensity	.407	.108	.000*	1.502	1.216	1.855
7	Intercept	1.388	.071	.000			
	Therapy intensity	.189	.117	.106	1.208	.961	1.520
8	Intercept	.581	.063	.000			
	Therapy intensity	.220	.079	.006*	1.246	1.066	1.456

*odds ratio is significant to p<0.05

Table K3

Logistic regression output for mBCH profiles and therapeutic intensity as predictors of reliable recovery, split by profile

						95% Confidence interval for odds ratio	
Profile	**Variable**	**Beta**	**Standard error**	**Significance**	**Odds ratio**	**Lower bound**	**Upper bound**
1	Intercept	.014	.117	.907			
	Therapy intensity	.209	.141	.136	1.233	.936	1.624
2	Intercept	-.137	.067	.041			
	Therapy intensity	.319	.079	.000*	1.376	1.178	1.608
3	Intercept	.312	.240	.194			
	Therapy intensity	.359	.351	.307	1.432	.719	2.850
4	Intercept	.095	.126	.450			
	Therapy intensity	.041	.158	.797	1.041	.764	1.420
5	Intercept	.527	.087	.000			
	Therapy intensity	.175	.112	.118	1.191	.957	1.482
6	Intercept	.387	.074	.000			
	Therapy intensity	.355	.105	.001*	1.426	1.162	1.750
7	Intercept	1.376	.072	.000			
	Therapy intensity	.120	.117	.305	1.127	.897	1.417
8	Intercept	.639	.062	.000			
	Therapy intensity	.174	.079	.027*	1.191	1.020	1.389

*odds ratio is significant to p<0.05

Table K4

Logistic regression output for SP profiles and therapeutic intensity as predictors of reliable recovery, split by profile

						95% Confidence interval for odds ratio	
Profile	**Variable**	**Beta**	**Standard error**	**Significance**	**Odds ratio**	**Lower bound**	**Upper bound**
1.0	Intercept	.405	.264	.124			
	Therapy intensity	-.420	.313	.180	.657	.356	1.215
1.2	Intercept	-.064	.160	.690			
	Therapy intensity	.402	.192	.037*	1.495	1.025	2.180
1.6	Intercept	-.087	.295	.768			
	Therapy intensity	.671	.375	.074	1.956	.937	4.082
2.0	Intercept	-.176	.085	.038			
	Therapy intensity	.320	.101	.001*	1.377	1.131	1.677
2.6	Intercept	.106	.090	.241			
	Therapy intensity	.193	.109	.078	1.213	.978	1.503
3.0	Intercept	.847	.690	.220			
	Therapy intensity	-.714	.863	.408	.490	.090	2.656
3.4	Intercept	-.118	.344	.732			
	Therapy intensity	.057	.489	.907	1.059	.406	2.762
4.0	Intercept	.164	.120	.171			
	Therapy intensity	-.064	.154	.678	.938	.693	1.269
5.0	Intercept	.448	.101	.000			
	Therapy intensity	.013	.126	.920	1.013	.791	1.297
5.7	Intercept	1.118	.198	.000			
	Therapy intensity	.190	.285	.505	1.210	.692	2.116
6.0	Intercept	.291	.082	.000			
	Therapy intensity	.439	.112	.000*	1.552	1.246	1.932
7.0	Intercept	1.408	.072	.000			
	Therapy intensity	.324	.121	.007*	1.382	1.091	1.751
8.0	Intercept	.590	.061	.000			
	Therapy intensity	.197	.078	.011*	1.218	1.045	1.419

*odds ratio is significant to p<0.05

Appendix L : Assumptions used to define treatment recommendations

Table L1

Assumptions leading to treatment recommendations

Contextual consideration	Indication for treatment
IAPT is subject to the National Institute for Health and Clinical Excellence (NICE) guidance on stepped care, which states that "the least intrusive, most effective intervention is provided first" (NICE, 2009)	Low-intensity therapy is on average shorter and less intrusive than high-intensity, therefore should be offered by default unless there is no positive indication for high-intensity
NICE guidelines state "treatment should always have the best chance of delivering positive outcomes [for the client]" (National Health Service England, 2018)	Treatment should be recommended when there is a reasonable chance of benefit to the client. Benefit is assumed here to be ≥25% chance of reliable recovery or reliable improvement, following Saunders et al. (2016)*
IAPT should "act on non-improvement to enable stepping up to more intensive treatments, stepping down [to] a less intensive treatment...and stepping out when an alternative treatment or no treatment becomes appropriate" (NHS England, 2018)	Further treatment of a particular intensity should be recommended only when there is reasonable likelihood of improvement. Receiving any IAPT treatment should be recommended against when there is little indication of benefit to the client, i.e. where the chance of recovery or reliable change is <25%
All IAPT services are to aim for a recovery standard of at least 50% (NHS England, 2018)	The treatment intensity that offers a probability of recovery closest to or exceeding 50% should be offered. Where both intensities offer this, the least intrusive one should be prioritised

*The argument could be made that only recovery should be included under the definition of 'benefit', as clients whose symptoms improve but do not recover may be more likely to be re-referred to IAPT and repeat the cycle. However, as the NICE guidelines are broad in their definition of 'positive outcomes', I will be so as well.

Appendix M: Analysis of stepping treatment intensity up or down in ML and MAP profiles

In order to see whether 'stepping up' a client to high-intensity therapy following a course of low-intensity was useful, a logistic regression analysis was performed for clients who completed treatment, and received either low-intensity therapy (n=5735) or stepped up therapy (n=1614). Two ML profiles were significantly more likely to meet criteria for recovery or reliable change following a step up: Profile 1 (OR 1.5, 95%CI 1.2-2.0) and Profile 6 (OR 1.6, 95%CI 1.1-2.4). The same analysis was run for MAP profile 7 only, as this the only profile where outcomes varied between this study and Saunders' (2016). This profile was significantly more likely to achieve recovery or reliable change if stepped up (n=59) versus low-intensity treatment only (n=51); OR 1.5, 95%CI 1.1-2.1, $p<0.05$.

Repeating the analysis for 'stepped down' ML profiles (n = 220) compared to those receiving high-intensity therapy (n=3743) produced no significant odds ratios, suggesting clients who were stepped down were no more likely to recover compared to those who received an high-intensity intervention only. The same result was apparent for stepped-down MAP profile 7s (n=5) in relation to high-intensity only (n=150), although the sample size was too small to consider this a reliable result.