Advancing Precision Treatment Selection for Post-traumatic Stress Disorder

James Tait

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

School of Psychology
Faculty of Science
The University of Sheffield

Submission Date December 2024

ABSTRACT

Post-traumatic stress disorder is a mental health condition that develops following exposure to one or more traumatic events. In line with clinical practice guidelines, National Health Service (NHS) Talking Therapies services deliver Trauma-focussed Cognitive Behavioural Therapy (Tf-CBT) and Eye Movement Desensitisation and Reprocessing (EMDR) for the treatment of PTSD. Despite being evidence-based, many patients do not respond to these treatments and rates of reliable improvement are lower for PTSD than other mental health problems. There is evidence to suggest that some patients are more likely to respond to one of these therapies than the other (i.e., Tf-CBT vs. EMDR), and it may be possible to identify patients' optimal treatment from their pre-treatment data using machine learning methods. This is known as precision treatment selection. This thesis investigated whether precision treatment selection could improve treatment outcomes for PTSD in NHS Talking Therapies. First, a systematic review of studies that applied machine learning methods to predict the outcome of psychological therapy for PTSD was conducted. This revealed significant limitations and omissions in the application and reporting of machine learning methods, and an almost complete lack of external validation of prediction models. Second, a previously published precision treatment selection model was externally validated in an independent sample of NHS trauma therapy cases. This found that the model did not generalise to new patients, potentially due to methodological limitations. Third, a range of machine learning methods were applied to predict Tf-CBT outcomes using a large sample of NHS Talking Therapies trauma cases. Models were optimised and out-of-sample performance was compared in a validation sample, and methodological recommendations were made. It may be possible to develop a precision treatment selection model for PTSD, but this will require reliable application of machine learning methods in adequate clinical datasets.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisors, Jaime Delgadillo and Steve Kellett. I first met Jaime when I was an MSc student, at the time I was unsure what my next steps would be, but once I got to know Jaime and his research it was clear that the opportunity to take a PhD under his supervision was too good to miss. I continue to be inspired by Jaime's focus, ambition, and dedication to the highest standards of applied research, and perhaps even more so by his patience, generosity, and modesty. I met Steve on the first day of my PhD, and it was clear from the outset that Steve and Jaime complimented each other perfectly. Steve's pragmatic, clinician's-eye view has been invaluable throughout the process, as have his sense of humour and encouragement of interests outside of academia. I am immensely grateful for all of their support, guidance, and encouragement in completing my PhD, and in pursuing activities beyond my thesis, such as undertaking research visits, joining committees, and organising conference symposia, and I'm of course grateful for the chats over coffee (and occasionally curry) and spontaneous messages of praise and motivation. I can't imagine a better PhD supervisory team, and I can't thank them enough.

In our first supervision meeting Jaime and Steve emphasised the importance of research community. In this spirit, I would like to thank the Psychotherapy Evaluation and Research Lab at Sheffield (PEARLS) for the insightful lab meetings and sometimes even more insightful "pub-lab" that follows, the visits by inspiring researchers from around the globe, and visits to local restaurants. I am grateful to every member of PEARLS, but in particular I would like to thank Michael Barkham and Dave Saxon for sharing with me their data and innumerous pearls of wisdom.

I would like to thank Wolfgang Lutz, Ann-Kathrin Deisenhofer, Steffen Eberhardt, Brian Schwartz, and the rest of their research group at Trier, for hosting me at their lab and making such a tremendous effort to make that visit as enjoyable as it was intellectually fruitful. A true highlight.

I would like to thank Chris Gaskell and Taposhri Ganguly for making time to discuss research methods with me, Jennifer L. for providing second risk-of-bias ratings for my systematic review, and the Grounded Research team at RDaSH NHS Foundation Trust for their support.

I would like to thank the Society for Psychotherapy Research and all who are involved in the organisation for creating a such a welcoming international community dedicated to rigorous scientific research (and conviviality). The SPR conferences in Dublin and Ottawa were also highlights.

I would like to thank my thesis examiners, Nemanja Vaci and Nick Grey, for expressing their enthusiasm for the research in this thesis, for their comments which served to enrich the thesis, and for a thorough but rewarding viva.

Crucially, I would like to acknowledge the friends and companions that I have made along the way: James, Michael (Pete), Anton, Anna, Sonia, Dom, Catalina, Almudena, Ben, Fred, Myles, and many more, and of course, my fellow members of the Psychology PGR Society Committee: Eleanor, Billy, and Seb. The past three years would have been much more stressful and much less enjoyable without these people.

Finally, I would like to thank my family and friends outside of academia, especially my partner, Eimear, for their boundless patience, understanding and encouragement, keeping my stress levels in check and my life in perspective. I couldn’t have done it without you.

NOTES ON INCLUSION OF PUBLISHED WORK

The work presented in Chapter 3 was written up as a separate manuscript and accepted for publication. The co-authored paper is referenced below. This work is presented in Chapter 3 in a format that is cohesive with the thesis. Therefore, although there is some replication, Chapter 3 is not identical to the publication.

Tait, J., Kellett, S., Saxon, D., Deisenhofer, A.-K., Lutz, W., Barkham, M., & Delgadillo, J. (2024). Individual treatment selection for patients with post-traumatic stress disorder: External validation of a personalised advantage index. *Psychotherapy Research*, 1–14. https://doi.org/10.1080/10503307.2024.2360449

CONTENTS

LIST OF TABLES

LIST OF FIGURES

APPENDICES

CHAPTER 1
General Introduction

1.1 Post-traumatic Stress Disorder

The word *trauma* derives from the Greek word for *wound*, and has been part of medical nomenclature since the 17th century, where it is used to describe a sudden, severe physical injury, with an external cause, involving tissue damage and activation of the autonomic nervous system (Dumovich & Singh, 2024; Oxford University Press, 2023). Although chronic psychological reactions to acutely distressing experiences had been recorded and described in literature for millennia (Ben-Ezra, 2004), the scientific study of psychological trauma likely began in the 19th century. Eulenburg (1878) may have been the first to use the term *psychic trauma*, and did so to describe hypothesised neurological damage caused by an acutely distressing experience (Van Der Hart & Brown, 1990). By the end of the 19th century, *trauma* was commonly used as an analogy for psychological harm (Moskowitz et al., 2019). Notably, Breuer and Freud (1893) proposed that *psychical trauma*, arising from the interaction between distressing life events and the individual's psychological vulnerability, was a common cause of a diverse range of "hysterical symptoms". This included symptoms of *traumatic neuroses* (a term coined by Oppenheim, 1889), such as repeatedly "hallucinating" the distressing event.

Freud's work on the external origins of psychological distress culminated in *The aetiology of hysteria* (Freud, 1962; first published 1896), which presented a theory, based on clinical observation, that the common cause of "hysteria" was sexual abuse suffered in childhood, perpetrated by a caregiver or family member. Memories of the traumatic experience had been *repressed* and were not consciously accessible due to the overwhelming emotions they invoked. This theory was influenced by the early work of Janet, who proposed that diverse traumatic experiences can overwhelm psychological capacity leading to dissociation, which disrupts sensory and emotional processing and memory encoding in a way that has a lasting impact on behaviour and is resistant to therapeutic change (Van der Kolk & Van der Hart, 1989). However, Freud's (1896) theory was rejected by his peers (Lamprecht & Sack, 2002), and although congruent with modern theories of psychological trauma (Moskowitz et al., 2019), the work of Janet was largely forgotten for at least the first half of the 20th century (Ray, 2008; Van der Kolk & Van der Hart, 1989).

Nevertheless, the early 20th century saw increasing recognition that traumatic neuroses were frequently experienced by soldiers participating in combat. During the first world war soldiers were frequently diagnosed with conditions such as *soldiers' heart* (Myers, 1870), *effort syndrome* (Da Costa, 1871), and *shell shock* (Myers, 1915); symptoms of which included fatigue, palpitations, shortness of breath, tremors, headaches, vivid nightmares, and heightened startle response (Mott, 1919; Ray, 2008; Turnbull, 1998). Kardiner (1941) studied the long-term trauma neuroses experienced by first world war veterans and highlighted the interaction between psychological and physiological processes such as amnesia, "hallucinatory reproductions of sensations", heightened sensitivity to external stimuli, irritability, vivid nightmares, and sleep disturbances. This research influenced the inclusion of a diagnosis labelled *Gross stress reaction* in the first edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM; American Psychiatric Association, 1952). *Gross stress reaction* was listed under *Transient situational personality disorders* and was described as neurosis or psychosis experienced by a previously "normal personality" in response to overwhelming fear. The DSM did not describe specific symptoms, but claimed that *Gross stress reaction* was temporary, transient, and easily treated, and only occurred following military combat or civilian catastrophe. Contrary to this, Burgess and Holstrom (1974) observed that rape victims experienced nightmares and flashbacks comparable to those of war veterans.

Post-traumatic stress disorder (PTSD) was first operationally defined in the third edition of the DSM (DSM-III; American Psychiatric Association, 1980), influenced by Horowitz' (1976) comprehensive work on stress response syndromes, and advocates for Vietnam war veterans and victims of domestic and sexual abuse (Yehuda et al., 2015; Yehuda & McFarlane, 1995). Diagnostic criteria included "Existence of a recognizable stressor that would evoke significant symptoms of distress in almost everyone", reexperiencing symptoms such as intrusive memories or recurrent nightmares, emotional numbing or diminished interest, and at least two of the following: excessive autonomic arousal, sleep disturbance, survivor guilt, impaired memory or concentration, avoidance of reminders of the traumatic event, or intensification of symptoms in response to symbolic reminders of the event. Associated features included irritability, volatile anger, impulsive behaviour, and functional impairment, and dissociation was described as rare. Similar diagnostic criteria were subsequently included in *The ICD-10 Classification of Mental and Behavioural Disorders* (ICD-10; World Health Organization, 1993).

In the following decades these diagnostic manuals underwent numerous revisions (see Table 1.1 for the most recent versions of the DSM and ICD criteria for PTSD), but the PTSD diagnostic criteria remain somewhat unique in the prerequisite of exposure to an external traumatic event, whereas most DSM diagnoses are defined purely in terms of the combination of symptoms (Pai et al., 2017). However, the definition of a traumatic event is still the subject of much debate. Critics of the fourth edition of the DSM (DSM-IV; American Psychiatric Association, 1994) expressed concern that defining traumatic events in terms of the distress evoked in the individual was too open, as theoretically any experience could be considered traumatic, leading to "conceptual bracket creep" and overdiagnosis of PTSD (McNally, 2003;

Spitzer et al., 2007). Accordingly, the DSM-5 (DSM-5; American Psychiatric Association, 2013) removed any reference to subjective distress from the definition of trauma, and instead defines traumatic events as "Exposure to actual or threatened death, serious injury, or sexual violence". The ICD-11 criteria are more open to interpretation in that they include events of a "horrific nature", but specific examples of traumatic events are provided. Only a small proportion of people who experience these events develop PTSD (see section 1.2), suggesting that it is the individual's subjective response to an event that is traumatic, and not the event in and of itself (Ehlers & Clark, 2000; Moskowitz et al., 2019). Nevertheless, it is important to distinguish between post-traumatic stress and other forms of stress; a diverse range of stressful life experiences can contribute to the development of mental health problems (Kirkbride et al., 2024), but PTSD has a distinct psychobiological impact and important implications for treatment (Yehuda et al., 2015).

Table 1.1
Current Diagnostic Criteria for Post-traumatic Stress Disorder

Criterion	**DSM-5-TR (F43.10)**	**ICD-11 (6B40)**
Exposure to traumatic event	A. Exposure to actual or threatened death, serious injury, or sexual violence in one (or more) of the following ways: 1. Directly experiencing the traumatic event(s). 2. Witnessing, in person, the event(s) as it occurred to others. 3. Learning that the traumatic event(s) occurred to a close family member or close friend. In cases of actual or threatened death of a family member or friend, the event(s) must have been violent or accidental. 4. Experiencing repeated or extreme exposure to aversive details of the traumatic event(s) (e.g., first responders collecting human remains; police officers repeatedly exposed to details of child abuse).	Exposure to an event or situation (either short- or long-lasting) of an extremely threatening or horrific nature. Such events include, but are not limited to, directly experiencing natural or human-made disasters, combat, serious accidents, torture, sexual violence, terrorism, assault or acute life-threatening illness (e.g., a heart attack); witnessing the threatened or actual injury or death of others in a sudden, unexpected, or violent manner; and learning about the sudden, unexpected or violent death of a loved one.

	Note: Criterion A4 does not apply to exposure through electronic media, television, movies, or pictures, unless this exposure is work related.	
Intrusion/re-experiencing	B. Presence of one (or more) of the following intrusion symptoms associated with the traumatic event(s), beginning after the traumatic event(s) occurred: 1. Recurrent, involuntary, and intrusive distressing memories of the traumatic event(s). **Note:** In children older than 6 years, repetitive play may occur in which themes or aspects of the traumatic event(s) are expressed. 2. Recurrent distressing dreams in which the content and/or affect of the dream are related to the traumatic event(s). **Note:** In children, there may be frightening dreams without recognizable content. 3. Dissociative reactions (e.g., flashbacks) in which the individual feels or acts as if the traumatic event(s) were recurring. (Such reactions may occur on a continuum, with the most extreme expression being a complete loss of awareness of present surroundings.) **Note:** In children, trauma-specific reenactment may occur in play. 4. Intense or prolonged psychological distress at exposure to internal or external cues that symbolize or resemble an aspect of the traumatic event(s). 5. Marked physiological reactions to internal or external cues that symbolize or resemble an aspect of the traumatic event(s).	Re-experiencing the traumatic event in the present, in which the event(s) is not just remembered but is experienced as occurring again in the here and now. This typically occurs in the form of vivid intrusive memories or images; flashbacks, which can vary from mild (there is a transient sense of the event occurring again in the present) to severe (there is a complete loss of awareness of present surroundings), or repetitive dreams or nightmares that are thematically related to the traumatic event(s). Re-experiencing is typically accompanied by strong or overwhelming emotions, such as fear or horror, and strong physical sensations. Re-experiencing in the present can also involve feelings of being overwhelmed or immersed in the same intense emotions that were experienced during the traumatic event, without a prominent cognitive aspect, and may occur in response to reminders of the event. Reflecting on or ruminating about the event(s) and remembering the feelings that one experienced at that time are not sufficient to meet the re-experiencing requirement.
Avoidance	C. Persistent avoidance of stimuli associated with the traumatic event(s), beginning after the traumatic event(s) occurred, as evidenced by one or both of the following: 1. Avoidance of or efforts to avoid distressing memories, thoughts, or feelings about or closely	Deliberate avoidance of reminders likely to produce re-experiencing of the traumatic event(s). This may take the form either of active internal avoidance of thoughts and memories related to the event(s), or external avoidance of people, conversations, activities, or situations reminiscent of the event(s). In extreme

	associated with the traumatic event(s). 2. Avoidance of or efforts to avoid external reminders (people, places, conversations, activities, objects, situations) that arouse distressing memories, thoughts, or feelings about or closely associated with the traumatic event(s).	cases the person may change their environment (e.g., move to a different city or change jobs) to avoid reminders.
Negative cognitions and mood	D. Negative alterations in cognitions and mood associated with the traumatic event(s), beginning or worsening after the traumatic event(s) occurred, as evidenced by two (or more) of the following: 1. Inability to remember an important aspect of the traumatic event(s) (typically due to dissociative amnesia and not to other factors such as head injury, alcohol, or drugs). 2. Persistent and exaggerated negative beliefs or expectations about oneself, others, or the world (e.g., “I am bad,” “No one can be trusted,” “The world is completely dangerous,” “My whole nervous system is permanently ruined”). 3. Persistent, distorted cognitions about the cause or consequences of the traumatic event(s) that lead the individual to blame himself/herself or others. 4. Persistent negative emotional state (e.g., fear, horror, anger, guilt, or shame). 5. Markedly diminished interest or participation in significant activities. 6. Feelings of detachment or estrangement from others. 7. Persistent inability to experience positive emotions (e.g., inability to experience happiness, satisfaction, or loving feelings).	[see Additional Clinical Features]
Arousal/ hypervigilance	E. Marked alterations in arousal and reactivity associated with the traumatic event(s), beginning or worsening after the traumatic event(s) occurred, as evidenced by two (or more) of the following:	Persistent perceptions of heightened current threat, for example as indicated by hypervigilance or an enhanced startle reaction to stimuli such as unexpected noises. Hypervigilant persons constantly guard themselves against danger and feel

	1. Irritable behavior and angry outbursts (with little or no provocation) typically expressed as verbal or physical aggression toward people or objects. 2. Reckless or self-destructive behavior. 3. Hypervigilance. 4. Exaggerated startle response. 5. Problems with concentration. 6. Sleep disturbance (e.g., difficulty falling or staying asleep or restless sleep).	themselves or others close to them to be under immediate threat either in specific situations or more generally. They may adopt new behaviours designed to ensure safety (e.g., not sitting with ones' back to the door, repeated checking in vehicles' rear-view mirrors).
Duration	F. Duration of the disturbance (Criteria B, C, D, and E) is more than 1 month.	Following the traumatic event or situation, the development of a characteristic syndrome lasting for at least several weeks, consisting of all three core elements: [Re-experiencing, avoidance, and hypervigilance]
Functional impairment	G. The disturbance causes clinically significant distress or impairment in social, occupational, or other important areas of functioning.	The disturbance results in significant impairment in personal, family, social, educational, occupational or other important areas of functioning. If functioning is maintained, it is only through significant additional effort.
Absence of confounding factors	H. The disturbance is not attributable to the physiological effects of a substance (e.g., medication, alcohol) or another medical condition.	
Dissociative and other additional features	*Specify* whether: **With dissociative symptoms:** The individual's symptoms meet the criteria for posttraumatic stress disorder, and in addition, in response to the stressor, the individual experiences persistent or recurrent symptoms of either of the following: 1. **Depersonalization:** Persistent or recurrent experiences of feeling detached from, and as if one were an outside observer of, one's mental processes or body (e.g., feeling as though one were in a dream; feeling a sense of unreality of self or body or of time moving slowly). 2. **Derealization:** Persistent or recurrent experiences of unreality of surroundings (e.g., the world around the individual is experienced as unreal, dreamlike, distant, or distorted).	Additional Clinical Features: • Common symptomatic presentations of Post-Traumatic Stress Disorder may also include general dysphoria, dissociative symptoms, somatic complaints, suicidal ideation and behaviour, social withdrawal, excessive alcohol or drug use to avoid re-experiencing or manage emotional reactions, anxiety symptoms including panic, and obsessions or compulsions in response to memories or reminders of the trauma. • The emotional experience of individuals with Post-Traumatic Stress Disorder commonly includes anger, shame, sadness, humiliation, or guilt, including survivor guilt.

	Note: To use this subtype, the dissociative symptoms must not be attributable to the physiological effects of a substance (e.g., blackouts, behavior during alcohol intoxication) or another medical condition (e.g., complex partial seizures).	
Course	*Specify* if: **With delayed expression:** If the full diagnostic criteria are not met until at least 6 months after the event (although the onset and expression of some symptoms may be immediate).	Course Features: • Onset of Post-Traumatic Stress Disorder can occur at any time during the life span following exposure to a traumatic event. • Onset of Post-Traumatic Stress Disorder symptoms typically occurs within three months following exposure to a traumatic event. However, delays in the expression of Post-Traumatic Stress Disorder symptomology can occur even years after exposure to a traumatic event. • The symptoms and course of Post-Traumatic Stress Disorder can vary significantly over time and individuals. Recurrence of symptoms may occur after to exposure to reminders of the traumatic event or as a result of experiencing additional life stressors or traumatic events. Some individuals diagnosed with Post-Traumatic Stress Disorder can experience persistent symptoms for months or years without reprieve. • Nearly one half of individuals diagnosed with Post-Traumatic Stress Disorder will experience complete recovery of symptoms within 3 months of onset.

Note. DSM-5-TR = Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision (American Psychiatric Association, 2022); ICD-11 = International Classification of Diseases 11th Revision (World Health Organization, 2021).

1.2 Impact of PTSD

1.2.1 Prevalence of PTSD

Koenen et al. (2017) analysed data from 26 population surveys in the World Health Organization World Mental Health Surveys, collected between 2001 and 2012, in 24 countries ranging from low to high income, in Africa, America, Asia, Europe, and Oceania. England, Scotland, and Wales were not included, but the United Kingdom falls into the high income category (World Health Organization, 2024). Through diagnostic interviews, PTSD was assessed against DSM-IV criteria, and exposure to potentially traumatic events was measured across six categories: war-related, physical assault, sexual assault, threats to personal integrity, threats to loved ones, and the traumatic death of a loved one. The lifetime prevalence of PTSD (i.e., proportion of the sample diagnosed with PTSD at some point during their lifetime) across all countries combined was 3.9%, and within country lifetime prevalence ranged from 0.3% (China) to 8.8% (Northern Ireland). The rate of exposure to potentially traumatic events across all countries was 69.7% and ranged from 28.6% (Bulgaria) to 84.6% (Ukraine). Among those exposed to such events, lifetime prevalence of PTSD was 5.6% and ranged from 0.5% (China) to 14.5% (Northern Ireland); twelve-month prevalence of PTSD was 2.8% and ranged from 0.2% (Peru) to 8.4% (Northern Ireland); and one month prevalence of PTSD was 1.4% and ranged from 0.1% (China and Peru) to 4% (Northern Ireland).

Lifetime prevalence of PTSD was significantly higher in high income countries (5%) than low to lower-middle income (2.1%) and upper-middle income (2.3%) countries. However, across all countries low income was significantly associated with higher rate of exposure to potentially traumatic events, lifetime prevalence of PTSD, and persistent PTSD (lasting 12 months or more). Being of younger age, female, unemployed, unmarried, and less educated were associated with higher lifetime prevalence of PTSD among those exposed to potentially traumatic events, suggesting that people with these demographic characteristics are more likely to develop PTSD.

In England, The Adult Psychiatric Morbidity Survey 2014 (Fear et al., 2016) interviewed 7,546 adults (aged 16 and over) from the general population and found that 31.4% (95% CI [30.0%, 32.7%]) had experienced a potentially traumatic event at some point during their lifetime. Compared to the findings of Koenen et al. (2017), this is below the global average (69.7%), below the average for high income countries (72.4%), and below the lowest trauma exposure rate for both high income countries and Western European countries (Spain = 54%). However, the World Health Organization World Mental Health Survey definition of a potentially traumatic event was broader than that of The Adult Psychiatric Morbidity Survey, which did not include life-threatening illness in oneself or a loved one (in line with DSM-5 and ICD-11 criteria).

Nevertheless, Fear et al. (2016) found that 4.4% (95% CI [3.8%, 5%]) of adults in England screened positive for PTSD in the past month (using the self-report PTSD Checklist – Civilian version), 3.3% believed they had had PTSD at some point in their lives, and 1.9% had been diagnosed with PTSD by a professional. Younger age was associated with a higher positive screening rate, and there was an interaction between age and sex, in that the positive screening rate for women peaked in the 16-24 age bracket (12.6%) and then decreased rapidly with age, whereas for men the positive screening rate was between 3.6% and 5% across age groups until falling rapidly from 65 and above. The rate of exposure to potentially traumatic events did not vary by ethnic group, but the positive screening rate for Black/Black British respondents (8.3%) was almost double that of White British respondents (4.2%), and this approached statistical significance (95% confidence). Unemployment, receiving benefits, and living alone were associated with higher positive screening rates for PTSD. This is congruent with evidence that poverty and deprivation are associated with greater prevalence and severity of PTSD from cross-sectional studies in in England (Cowlishaw et al., 2021; Delgadillo & Richardson, 2024) and longitudinal cohort studies in the USA (Lowe et al., 2014; Ravi et al., 2023).

Additionally, systematic reviews have found that higher rates of trauma exposure and PTSD are reported amongst groups such as emergency service workers (Berger et al., 2012; Jones, 2017), prisoners (Baranyi et al., 2018), combat veterans (Hines et al., 2014; Ramchand et al., 2015), and refugees and asylum seekers (Blackmore et al., 2020).

1.2.1.1 Rise in Prevalence During the COVID-19 Pandemic

Shevlin et al. (2020) surveyed a representative sample of 2,025 participants across the UK during the first week of social restrictions in response to the COVID-19 pandemic (“lockdown”) in March 2020 and found that 16.79% of the sample met ICD-11 criteria for PTSD (95% CI [15.2%, 18.4%]). This is considerably higher than the earlier estimates for England (Fear et al., 2016) and Northern Ireland (Koenen et al., 2017) described above. However, Shevlin et al. (2020) advise caution in making these comparisons as it is unclear whether the COVID-19 pandemic qualifies as a traumatic event. In a reverse of the findings of Fear et al. (2016) and Koenen et al. (2017), Shevlin et al. (2020) found that during lockdown, PTSD prevalence was significantly higher for males (18.9%) than females (14.9%). Living in a house with children, and a higher perception of the risk of COVID-19 infection were also associated with higher prevalence of PTSD. Concurrently, a meta-analysis of data from 24 countries (Yunitri et al., 2022) found a PTSD prevalence rate of 17.52% (95% CI [13.89%, 21.86%]) during the COVID-19 pandemic, with higher rates among those living in European countries (25.05%), working in COVID-19 units (30.98%), and nurses (28.22%). Although, the rate also varied significantly depending on which measure of PTSD a study used.

1.2.2 Health Impact of PTSD

PTSD is typically chronic, the World Health Organization World Mental Health Surveys found that the mean duration was 6 years, and this rose to 13 years for combat-related PTSD (Shalev et al., 2017). PTSD is associated with poor health outcomes including cardio-respiratory health problems, gastro-intestinal health problems, chronic pain, diseases of the bones and joints, autoimmune disorders, dementia, and psychosocial outcomes such as poorer quality of life, disability, and increased risk of suicide (O'Donovan et al., 2015; Pacella et al., 2013; Sareen et al., 2007). PTSD often co-occurs with other mental health problems such as alcohol use disorder (Debell et al., 2014), borderline personality disorder (Knefel et al., 2016; Scheiderer et al., 2015), depression, anxiety, suicidality, and self-harm (Karatzias et al., 2019; Spinhoven et al., 2014). Fear et al. (2016) found that among those who screened positive for PTSD in England in 2014, 50.9% were receiving treatment for mental health problems, including 24% who were receiving psychological therapy.

1.3 Treatment of PTSD

1.3.1 Psychological Therapies for Post-traumatic Stress Disorder

Current (i.e., updated in the past five years; Shekelle et al., 2001) Clinical Practice Guidelines (CPG) recommend a minimum of eight sessions of individual, manualised, trauma-focussed psychological therapy as the first-line treatment for PTSD in adults. CPG are intended to bridge the gap between treatment research and delivery. In a systematic review of English language CPG for PTSD, Martin et al. (2021) rated the quality of fourteen CPG using a validated measure for the evaluation of health-care guidelines. Six CPG were recommended for use, and four of these were published or updated within the five years preceding the start of this programme of PhD research: American Psychological Association (2017), National Institute for Health and Care Excellence (2018), Phoenix Australia Centre for Posttraumatic Mental Health (2021), and Veterans Affairs/Department of Defence (2017). Additionally, Martin et al. (2021) note that the International Society for Traumatic Stress Studies (ISTSS; 2018) guideline may have been recommended for use if the evidence summary documents and reference lists were publicly available. Table 1.2 presents the psychological therapies that received a strong recommendation in at least one of these CPG. Trauma-focussed Cognitive Behavioural Therapy (Tf-CBT) is an umbrella term that can include cognitive processing therapy (CPT; Resick et al., 2017), prolonged exposure (PE; Foa et al., 2019), and cognitive therapy for PTSD (CT-PTSD; Ehlers & Wild, 2020). Tf-CBT are the most widely recommended treatments across CPG, along with eye movement desensitization and reprocessing (EMDR; Shapiro, 2018). The underlying theory and core techniques of these therapies are briefly described in Appendix A.

Table 1.2
Psychological Therapies Recommended for Treatment of Post-traumatic Stress Disorder by Clinical Practice Guidelines

Treatment	Clinical Practice Guideline				
	APA[1]	ISTSS[2]	NICE[3]	PAC[4]	Va/DoD[1]
Trauma focussed-CBT[5]	✓✓[6]	✓✓	✓✓	✓✓	
Cognitive Processing Therapy	✓✓	✓✓	✓✓	✓✓	✓✓
Cognitive Therapy for PTSD			✓✓		✓
CBT without a trauma focus	✓✓[6]	✓			
Cognitive Therapy	✓✓[6]	✓✓		✓✓	
EMDR	✓	✓✓	✓✓[7]	✓✓	✓✓
Prolonged Exposure	✓✓	✓✓	✓✓	✓✓	✓✓
Narrative Exposure Therapy	✓	✓	✓✓	✓	

Note. APA = American Psychological Association (2017); ISTSS = International Society for Traumatic Stress Studies (2018); NICE = National Institute for Health and Care Excellence (2018b); PAC = Phoenix Australia Centre for Posttraumatic Mental Health (2021); Va/DoD = Veterans Affairs/Department of Defence (2023).

✓✓ = Strong recommendation.

✓ = Weak/conditional recommendation.

[1] USA

[2] International

[3] UK

[4] Australia

[5] Trauma focussed-CBT is an umbrella term that encompasses numerous variants, including Cognitive Processing Therapy, Cognitive Therapy for PTSD, and Prolonged Exposure.

[6] The APA (2017) guideline does not distinguish between Trauma focussed-CBT and other forms of CBT, nor between Cognitive Therapy for PTSD and other forms of cognitive therapy.

[7] NICE (2018b) guideline recommends offering EMDR to individuals presenting with non-combat-related trauma.

CPG recommendations are based on systematic reviews and meta-analyses of randomised controlled trials (RCT). A recent network meta-analysis by Merz et al. (2019) found that psychological therapies for PTSD are at least as effective as pharmacotherapy in the short-term, and more effective long-term. This is supported by an earlier systematic review, which found larger effect sizes for trauma-focussed psychological interventions than pharmacological interventions when each were compared to active control/placebo conditions (Lee et al., 2016). Additionally, there is evidence that a majority of patients prefer psychological therapy (including trauma-focussed therapy) to pharmacological therapy (Simiola et al., 2015; Swift et al., 2017), and patient preference is related to better treatment outcomes (Zoellner et al., 2019).

In systematic reviews and meta-analyses of RCT comparing psychological therapies for PTSD, Lewis et al. (2020) found that EMDR and Tf-CBT (in particular CPT, CT-PTSD, and PE) had the strongest effects, and Mavranezouli et al. (2020) found that the effects of Tf-CBT and EMDR were maintained at 1-4 month follow up. In a network meta-analysis comparing different forms of Tf-CBT, Jericho et al. (2021) found that CPT was the most effective. This review influenced the VA/DoD decision to downgrade CT-PTSD from a strong recommendation to a weak recommendation in the recent update to their CPG for PTSD (VA/DoD, 2023). Jericho et al. (2021) also reported that meta-cognitive therapy (Wells & Sembi, 2004) was superior to other forms of Tf-CBT, but this meta-analysis only included two RCT of meta-cognitive therapy with 32 and 20 participants respectively (Wells et al., 2015; Wells & Colbear, 2012).

1.3.2 Treatment of PTSD in the NHS Talking Therapies Programme

The National Health Service (NHS) in England began offering National Institute for Health and Care Excellence (NICE) recommended psychological therapies in primary care settings in 2008 via the NHS Talking Therapies programme (Clark et al., 2009). NHS Talking Therapies (formerly known as Improving Access to Psychological Therapies, IAPT; Clark & Whittington, 2023) implements a stepped care model. Therefore, most patients are initially offered low intensity interventions (e.g., CBT based self-help and psychoeducational peer-support groups) and those who do not respond can be subsequently “stepped-up” to high intensity interventions (i.e., individual psychological therapy, mainly CBT). Low intensity interventions are not recommended for PTSD due to lack of evidence for their effectiveness to treat this condition, and patients with PTSD are assigned directly to high intensity interventions (i.e., Tf-CBT or EMDR) at screening. High intensity therapists are qualified to a postgraduate level and practice under regular clinical supervision by experienced therapists. High intensity therapists receive training in Tf-CBT (CT-PTSD/CPT/PE) as part of their core CBT training (Hool, 2010), and can opt to train in EMDR following completion of core training and two years’ experience delivering psychological therapies (Health Education England, 2021). Patients either self-refer to NHS Talking Therapies services or are referred by their general practitioner. Patients are not routinely excluded due to current drug or alcohol misuse, or concurrent mental health problems that require secondary mental healthcare (e.g., psychosis, bipolar disorder, personality disorders, or eating disorders); patients are assessed on an individual basis and may be referred to more intensive, multi-professional care where necessary (National Collaborating Centre for Mental Health, 2018). In the year 2022-2023, over 1.2 million patients accessed treatment at NHS Talking Therapies services, including over 66,000 patients with PTSD (NHS Digital, 2024).

1.3.3 Treatment Response and Acceptability

Although average treatment effect sizes for the most empirically supported psychological therapies for PTSD are medium to large (Jericho et al., 2021), a significant proportion of patients do not respond to treatment. The recent systematic reviews described above did not report treatment response rates, but in an earlier systematic review, Schottenbauer et al. (2008) found that non-response rates ranged from 20%-67% for PE, 3.6%-48% for CPT, and 7.3%-92% for EMDR; although, the definition of “non-response” may vary across the studies reviewed. In a smaller but more recent review of treatment for combat-related PTSD, Steenkamp et al. (2015) found that 60%-72% of patients still met diagnostic criteria for PTSD after receiving CPT or PE. Response rates may be even lower in routine clinical practice than in RCTs. An analysis of patient records from 16 NHS Talking Therapies services by Robinson et al. (2020) found that only 32% of 2,493 patients accessing Tf-CBT for PTSD attained a *reliable and clinically significant improvement in symptoms* (indicated by a decrease in symptom score, equal to or greater than a statistically derived reliable change index, to a score below an established diagnostic cut-off; Jacobson & Truax, 1991). This is well below the average treatment response rate across all NHS Talking Therapies patients, which is around 50% (NHS Digital, 2024).

A contributing factor to nonresponse is early termination of treatment by the patient, or *dropout*, potentially due to poor acceptability of the treatment approach. Robinson et al. (2020) estimated that between 6 and 16 sessions of Tf-CBT are required for reliable and clinically significant improvement in PTSD symptoms, but the median treatment length was 6 sessions. Lewis et al. (2020) systematically reviewed dropout from RCTs of psychological therapies for PTSD, and found that the pooled dropout rate was 16% (95% CI [14, 18%]), suggesting that around one in six patients dropout. Furthermore, although demographic and trauma characteristics were not associated with dropout, the dropout rate was significantly higher for trauma focussed therapies than for non-trauma focussed therapies. The pooled dropout rate for EMDR was 18% (95% CI [12, 24%]), for PE was 22% (95% CI [16%, 28%]), and for CPT was 30% (95% CI [22%, 39%]). This indicates that patients with PTSD are most likely to drop out from the treatments that are most efficacious on average, and dropout rates may be even higher in routine clinical practice than in RCTs (Najavits, 2015).

1.3.4 Complex PTSD

An expert consensus statement published by the International Society of Traumatic Stress Studies (Cloitre et al., 2012) recommended a distinct, multi-phase treatment approach for patients diagnosed with Complex PTSD (CPTSD). The concept of CPTSD emerged due to concern that the current PTSD criteria did not account for the full range of traumatic experiences and posttraumatic stress responses, in particular responses to repeated or prolonged trauma (Herman, 1992; Maercker, 2021). The ICD-11 diagnostic criteria for CPTSD (WHO, 2024) include the core PTSD symptoms of reexperiencing, avoidance, and hyperarousal, with the addition of emotion regulation difficulties, relationship difficulties, and negative self-concept. Accordingly, the proposed multi-phase treatment approach begins with a “stabilisation” phase, designed to teach self-regulation, before progressing to trauma-focussed therapy. However, there has been debate as to whether PTSD and CPTSD are distinct conditions, and whether the distinction is clinically useful (Knefel et al., 2016; Resick et al., 2012). Unlike the ICD-11, the DSM-5 does not include separate diagnostic criteria for CPTSD, and there is considerable overlap between the ICD-11 diagnostic criteria for CPTSD the DSM-5 diagnostic criteria for PTSD (see Table 1.1). A review of empirical evidence found that the stabilisation treatment phase did not improve treatment outcomes for patients with CPTSD, and that this approach potentially delayed or prevented access to effective treatment (De Jongh et al., 2016). This is supported by recent evidence from clinical trials, which found that CPTSD was not associated with poorer treatment outcomes in response to trauma-focussed therapies (Bækkelund et al., 2022; Voorendonk et al., 2020), and CPTSD treatment outcomes were not improved by integrating stabilisation treatment (Hoeboer et al., 2021). As such, NHS Talking Therapies do not currently distinguish between PTSD and CPTSD when offering treatments to patients, and no such distinction will be made in the following chapters of this thesis.

1.4 Precision Treatment Selection

One way that PTSD treatment outcomes might be improved, is by developing new, more effective psychological therapies. However, as described above, a multitude of psychological therapies for PTSD have been developed in the past 50 years, and those with the strongest evidence base are roughly equally efficacious. This is consistent with the *Dodo bird hypothesis* proposed by Rosenzweig (1936) and supported by empirical evidence since the 1970s (Luborsky et al., 1975; Smith & Glass, 1977). The Dodo bird hypothesis refers to the observation that all effective variants of psychological therapy are equally effective on average, which may suggest that it is their *common factors* that are important for therapeutic change, rather than their distinguishing features (Rosenzweig, 1936; Wampold, 2015). To some extent, treatments for PTSD are an exception to this, as meta-analyses consistently find that that trauma-focussed psychological therapies are superior to therapies without a trauma focus (Lewis et al. 2020; Mavranezouli et al., 2020). However, there are many different evidence-based trauma-focussed therapies, and it could be argued that it is the elements shared by different trauma-focussed psychological therapies that are important for therapeutic change, rather than any element unique to one particular approach (Wampold, 2019). Nevertheless, comparing the average effectiveness of treatments does not account for individual differences in treatment response. PTSD is a complex, heterogenous condition (Galatzer-Levy & Bryant, 2013), and there is significant variability in response to treatment for PTSD, suggesting that although different treatments may be equally effective on average, different patients do not respond to them equally (Herzog & Kaiser, 2022).

Hence, another way that PTSD treatment outcomes might be improved is through *personalised treatment selection*. This involves tailoring treatment delivery to the individual by selecting the most appropriate treatment type, intensity, or components, for a specific patient with specific problems (Cohen et al., 2021). To some extent, all psychological therapies involve a degree of personalisation, such as case formulation in cognitive behavioural therapies (Persons, 2022), and the choice of treatment type is often based on CPG, clinical intuition, and patient preference. However, the patient characteristics that are associated with treatment response are numerous, interrelated, and complex (Delgadillo et al., 2017); therefore, predicting the optimal course of treatment for each individual patient is a challenging task. The question of *what works for whom* has occupied psychological therapy researchers for over fifty years (Paul, 1967), and in that time research has consistently shown that judgements made by statistical algorithms are typically more accurate than those made by expert clinicians (Ægisdóttir et al., 2006; Meehl, 1954). The use of data driven methods to optimally personalise treatment can be referred to as *precision* treatment selection (Deisenhofer et al., 2024; DeRubeis, 2019).

Historically, this was conceptualised as *aptitude-by-treatment interactions* (ATI), whereby individual patient characteristics interact with different psychological therapy approaches, techniques, or mechanisms of action to moderate treatment effects (Cronbach & Snow, 1977). Variables that moderate treatment effects in this way can be referred to as *prescriptive* predictor variables, as opposed to *prognostic* predictor variables that predict outcome regardless of treatment type (Cohen & DeRubeis, 2018). However, reliably detecting interaction effects requires substantial statistical power, most psychological therapy study samples are underpowered, and ATI effects are small and numerous. Systematic reviews have found that a large number of variables each explain a small proportion of variance in PTSD treatment outcome (Barawi et al., 2020; Dewar et al., 2020; Malejko et al., 2017), and it is likely that many of these variables covary, interact, or are non-linearly related. The ATI approach was therefore deemed to have been of little practical clinical utility (DeRubeis, 2019; Kessler et al., 2017; Snow, 1991). However, recent advances in computing power and availability of large psychological therapy datasets have rekindled interest in precision treatment selection, as many psychological therapy researchers have begun to utilise Machine Learning (ML) methods, which may be well suited to the task (Aafjes-van Doorn et al., 2021; Chekroud et al., 2021).

1.4.1 Machine Learning Methods

Broadly, there are two approaches to statistical modelling. The first seeks to *explain* relationships between variables by fitting a model to a dataset, evaluating model fit, and thereby making inferences about the process or mechanism that explains such associations. This *explanatory* approach broadly underpins many conventional statistical analyses applied in the sciences, and it follows a hypothesis-testing approach, where expected relationships are specified a priori. The second, *algorithmic* approach, seeks to discover patterns in available data, without prior specification of expected relationships, and with the practical goal of solving prediction and/or classification tasks. The latter approach assumes that the relationships between variables are complex and at least partly unknowable (Breiman, 2001b). This distinction is often overlooked in psychological research, with evidence for relationships between variables in the current dataset interpreted as evidence of prediction (Yarkoni & Westfall, 2017).

Machine Learning (ML) refers to a family of statistical methods that developed in the field of computer science, which follow the algorithmic approach to data analysis (Dwyer et al., 2018). ML methods use *algorithms* to detect (or *learn*) patterns in data, which can be used to develop *models* that make predictions in new data. An algorithm is a set of mathematical processes performed by a computer to solve a particular problem, and a model is a mathematical representation of the relationships between a set of variables. ML methods can be divided into two broad categories: *supervised* and *unsupervised*. In supervised ML an outcome variable is specified, and the algorithm aims to identify predictors of the outcome variable. In unsupervised ML, an outcome variable is not specified, and the algorithm aims to classify cases into groups with similar combinations of characteristics (Hastie et al., 2009). In the field of ML, explaining the mechanisms or processes underlying the associations between variables is not always possible nor necessary, as the priority in this algorithmic approach to data analysis is to maximise accuracy in prediction or classification tasks (Yarkoni & Westfall, 2017).

1.4.1.1 The Machine Learning Pipeline

When applied optimally, the ML approach to prediction modelling follows a sequence of six steps that can be referred to as the *ML pipeline* (depicted in Figure 1.1; Delgadillo & Atzil-Slonim, 2022). In step one, a sample size calculation is performed to determine the required sample size, considering the specific ML method, the number of candidate predictor variable parameters to be estimated, and the expected performance of the model.

In step two, any necessary *pre-processing* of the data is performed. This can include imputation of missing data, reduction of categorical variables, steps to mitigate the effect of class imbalance, and case-control matching to account for non-random allocation to conditions.

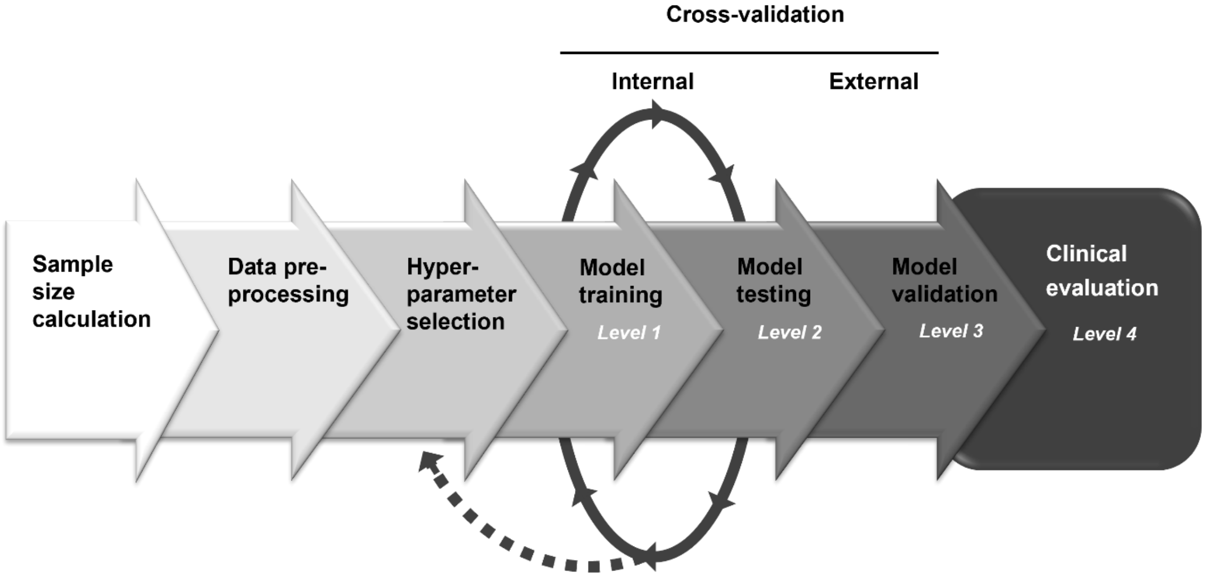
In step three, *hyperparameter selection* takes place. Hyperparameters are parameters that are set by the user that influence how an algorithm arrives at a solution. Hyperparameter values can be selected a priori based on previous empirical evidence or theory, or they can be *tuned* on the current dataset to identify the optimal values. However, if hyperparameter tuning is performed manually by the researcher this can lead to *overfitting* (i.e., capitalising on the idiosyncrasies of the training data to the detriment of generalisability), therefore it is important to apply internal cross-validation when hyperparameter tuning, in a process known as *grid search optimisation*.

In step four, the ML model is developed using a training dataset, this can include selection/exclusion of predictor variables (model specification), and estimation of model parameters (e.g., coefficients). If model performance is evaluated at this stage within the training dataset, then the extent that the model will generalise beyond the training dataset is unknown; this is often referred to as *apparent validation* and for the purposes of this thesis will be considered *level 1* evidence (i.e., evidence with a low level of reliability on a four-point scale of reliability; Delgadillo & Atzil-Slonim, 2022).

In step five, model performance is tested by applying some form of internal cross-validation (*level 2* evidence). A common example of this is *k*-fold cross-validation, whereby the training dataset is divided into a certain number (*k*, often 5 or 10) of subsamples (*folds*), and predictions for participants in each fold are made by training the model parameters on the data from participants in all other folds, repeated *k* times (*leave-one-out* is an extreme form of *k*-fold where $k = N$). The primary goal of internal-cross validation is to limit optimism in estimates of prediction accuracy and error by temporarily excluding each participant's data from the estimation of parameters when predicting their outcome.

In step six, the model is externally validated by applying the predictors selected and parameters estimated in the training data to predict outcomes for a statistically independent dataset and evaluating model performance (*level 3* evidence). The simplest method of external validation is to randomly split a dataset into training and *hold-out* validation data. However, if a hold-out dataset was collected in the same time period and location by the same researchers (as is often the case with randomly split datasets), then it is not completely independent; More stringent forms of external validation include *geographic* validation, where the model is tested on data from a different geographic location, and *temporal* validation, where the model is tested on data collected during a different time period (Steyerberg, 2019). Out-of-sample prediction accuracy and error can now be evaluated, and measures of calibration and discrimination help to assess the model's ability to make individual predictions.

Figure 1.1
The Machine Learning Pipeline



Note. Figure adapted from Delgadillo and Atzil-Slonim (2022).

1.4.1.2 External Validation

Before a prediction model can be evaluated in clinical practice, it first requires external validation to demonstrate that the model's predictive capabilities reliably generalise beyond the data that was used to develop, or *train,* the model. Prediction models are likely to make more accurate predictions in the data used to train the model, than in data previously unseen by the model (Siontis et al., 2015). In the worst case scenario, the model is *overfit* to the training data, and describes idiosyncratic relationships between variables that do not generalise to different samples drawn from the same population (Steyerberg, 2019).

External validation is closely related to the concept of replication; for an empirical finding to be credible, it must be repeatable, and the results consistent over repetitions with different samples from the same (or a related) population (Nosek et al., 2022). The importance of replication in psychological science was highlighted in the last decade by what some refer to as the *replication crisis* (Wiggins & Christopherson, 2019), during which a substantial proportion of attempts to replicate key psychological findings failed, suggesting that the initial findings were false positives (Open Science Collaboration, 2015). Despite this context and the credibility brought by external validation, applications of the method are still relatively rare. Researchers can develop new prediction models, but without external validation the models create little clinical traction and create research waste. Two recent systematic reviews of clinical prediction models in psychiatry found evidence of lack of external validation. Meehan et al. (2022) found that only 20.1% of 308 models were externally validated in an independent sample, whilst Salazar de Pablo et al. (2021) found that just 4.6% of 584 models were externally validated.

1.4.2 Precision Treatment Selection Using Machine Learning Methods

Building on earlier work by Barber and Muenz (1996), DeRubeis et al. (2014) developed a precision treatment selection method called the *Personalised Advantage Index* (PAI). The PAI method uses a statistical model to make a prediction about which of two alternative treatments may be most effective for each patient. The treatment with the best predicted outcome is labelled that patient's *optimal* treatment, and the other treatment is labelled their *suboptimal* treatment. Treatment outcomes are then retrospectively compared between patients who received their model-indicated optimal treatment, and patients who received their suboptimal treatment. Additionally, by subtracting the predicted outcome of one treatment from that of the other treatment, multiple patient characteristics can be reduced to a single continuous indicator of differential treatment response (i.e., the PAI score). The PAI score can be centred at zero, with negative values favouring one treatment option and positive values favouring the other treatment option (e.g., Delgadillo & Gonzalez Salas Duhne, 2020; Schwartz et al., 2021). A PAI close to zero would indicate that a patient is likely to respond similarly to either treatment option, but a PAI that is distant from zero would strongly favour a specific optimal treatment option. The early PAI studies (DeRubeis et al., 2014; Huibers et al., 2015) did not use ML methods, but, following the recommendation of Kessler et al. (2017), more recent PAI studies have begun to utilise ML methods due to their advantages when selecting predictors and estimating coefficients in ways that limit overfitting (e.g., Cohen et al., 2020; Delgadillo & Gonzalez Salas Duhne, 2020; Schwartz et al., 2021).

Three studies have tested the PAI approach for PTSD treatments delivered in the context of clinical trials and routine practice. Keefe et al. (2018) employed the PAI method to predict which treatment individual patients would be most likely to complete, using data from an RCT comparing CPT and PE for PTSD. This study found that patients who received their optimal treatment were significantly less likely to dropout (19.7% of patients who received their optimal treatment dropped out, compared to 40.5% of patients who received their suboptimal treatment). Hoeboer et al. (2021) used the PAI method to predict which treatment would yield the greatest change in symptoms for each patient in data from an RCT comparing PE with and without emotion regulation and interpersonal skills training. This study found that patients who received their optimal treatment had a significantly greater decrease in PTSD symptoms than patients who received their suboptimal treatment.

Deisenhofer et al. (2018) used the PAI method to predict whether patients with PTSD in a dataset of clinical case records from NHS Talking Therapies services were more likely to respond to Tf-CBT or EMDR in routine clinical practice. Using a genetic algorithm, Deisenhofer et al. (2018) developed two linear regression models to predict response to Tf-CBT and EMDR using participants' pre-treatment clinical and demographic characteristics. A genetic algorithm is a machine-learning optimisation algorithm that mimics Darwinian evolutionary processes (natural selection, cross-over, mutation) to build the best model from the available predictors (Mitchell, 1998). Deisenhofer et al. (2018) found a significantly higher rate of reliable improvement among patients who received their model-indicated optimal treatment (62.9% of patients who received their optimal treatment attained reliable improvement, compared to 33.6% of patients who received their suboptimal treatment). This suggests that by using a PAI it may be possible to predict the optimal treatment for individual patients with PTSD at the point of initial assessment in NHS Talking Therapies, and that applying treatment recommendations based on those predictions could significantly improve treatment outcomes for PTSD in routine practice.

However, although all three of these studies applied some form of internal cross-validation, none attempted to externally validate the PAI. Deisenhofer et al. (2018) used *leave-one-out* cross-validation, whereby each individual patient's outcome is predicted by fitting the regression model to the rest of the sample ($N - 1$), temporarily excluding that patient's data from the training sample. This adjusts for optimism in the estimation of prediction model performance, but it is possible that the model predictor selection and parameter estimation were biased towards the specific characteristics of the training sample, and for this reason the model predictions may not generalise to new data (Kessler et al., 2017; Steyerberg, 2019). External validation tests this by applying the same model to make predictions in new and separate outcome data and evaluating the accuracy of those predictions.

Thus far, much of the research applying ML methods to predict psychological therapy outcomes has focussed on the treatment of depression and anxiety, and relatively little has focussed on treatment for PTSD (Aafjes-van Doorn et al., 2021; Lee et al., 2018; Sajjadian et al., 2021). Vieira et al. (2022) systematically reviewed studies that applied ML methods to predict outcomes for CBT, but only included classification models (response vs. non-response), excluded other trauma-focussed psychological therapies (e.g., EMDR), and only included one study that sought to predict CBT outcomes in adults with PTSD. Ramos-Lima et al. (2020) systematically reviewed the use of ML methods in PTSD research but focussed on studies that sought to predict the presence or onset of PTSD and did not include any studies that sought to predict the outcome of CPG recommended psychological therapy for PTSD.

1.5 Overall Aims and Specific Objectives of the Thesis

The aim of this PhD thesis is to advance precision treatment for PTSD in NHS Talking Therapies services. This is potentially of benefit both to services and patients seeking treatment for PTSD. Given the recent interest in the potential for machine learning methods to advance precision psychological treatment, this thesis will rigorously explore this potential through the following objectives:

- Chapter 2 will systematically review studies that used machine learning methods to predict the outcome of psychological therapies for PTSD.
- Chapter 3 tests the external validity of a personalised advantage index for psychological therapies for PTSD developed to guide treatment selection in NHS Talking Therapies services.
- Chapter 4 will compare different machine learning methods at the task of predicting psychological therapy outcomes for PTSD, following best practice guidelines, and will investigate the effect of training sample size, in order to arrive at some recommendations for developing PTSD outcome prediction models.
- Chapter 5 will evaluate the state of precision treatment for PTSD in light of the findings presented in Chapters 2, 3, and 4, and the wider literature, and present recommendations for practice and research.

CHAPTER 2

Using Machine Learning Methods to Predict the Outcome of Psychological Therapies for Post-traumatic Stress Disorder: A Systematic Review

2.1 Introduction

This chapter aimed to conduct the first systematic review of studies that used machine learning (ML) methods to predict psychological therapy outcomes for post-traumatic stress disorder (PTSD). Given that this is an emerging literature, the focus of this review was on the application and reporting of each study's methods, benchmarked against the ML pipeline (Delgadillo & Atzil-Slonim, 2022) described in the previous chapter. The review question was framed following the recommendations of Moons et al. (2014) and Palazón-Bru et al. (2020) for framing systematic reviews of prognostic modelling studies, and the review was reported following PRISMA guidelines (Page et al., 2021).

2.2 Method

2.2.1 Pre-registration

The systematic review protocol was pre-registered with the PROSPERO database prior to conducting searches (Reference: CRD42022325021). The pre-registration can be accessed here:

https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42022325021

2.2.2 Eligibility Criteria

Inclusion and exclusion criteria are described in Table 2.1. To be included a study must have applied ML methods to pretreatment data to predict the outcome of a psychological therapy recommended by current clinical practice guidelines (CPG) as a first line treatment for current PTSD in adults. The inclusion criteria for this systematic review were guided by CPG that were based on systematic reviews, had been appraised to meet an acceptable quality standard using a standardised measure (Martin et al., 2021), and were published in the previous five years to ensure that they were up to date (Shekelle et al., 2001). This included the following CPG: American Psychological Association (2017), International Society for Traumatic Stress Studies (2018), National Institute for Health and Care Excellence (2018), Phoenix Australia Centre for Posttraumatic Mental Health (2021), and Veterans Affairs/Department of Defence (2017). The psychological therapies they recommend are presented in Table 1.2, and were predominantly trauma-focussed cognitive behavioural therapies, exposure-based therapies, and eye movement desensitisation and reprocessing.

Table 2.1
Systematic Review Inclusion and Exclusion Criteria

	Inclusion Criteria	Exclusion Criteria
Population	Adults (aged 18 and over) who received clinical practice guideline recommended psychological therapy for current PTSD.	Children and adolescents under the age of 18. Adults receiving treatment for a condition other than PTSD.
Intervention	Evidence-based psychological therapies recommended for the treatment of current symptoms of PTSD in adults by current clinical practice guidelines.	Psychological therapy intended to treat a different condition. Psychological therapy intended to prevent the onset or relapse of PTSD. Pharmacological therapy. Non-psychological therapy (e.g., acupuncture or yoga). Psychological therapy not

		recommended by clinical practice guidelines. (If any of the above were delivered alongside or in comparison to an intervention that met the inclusion criteria then that study would be included.)
Outcome to be predicted	Continuous or categorical outcomes of psychological therapy for PTSD, including remission, change in symptoms, dropout, and retention.	Future onset or relapse of PTSD. Current presence (diagnosis) of PTSD.
Time span of prediction	From pre-treatment to post-treatment. The outcome timepoint of interest is the end of treatment, or the follow-up nearest to the end of treatment.	
Intended moment of model use	Initial patient assessment, prior to the start of treatment.	During or after treatment.
Modelling approach	Prognostic models that applied supervised or unsupervised machine learning methods in the prediction of treatment outcomes from patients' pre-treatment or baseline features.	Diagnostic models that predict the presence of PTSD. Prognostic models that predict onset of PTSD. Modelling approaches that did not use any machine-learning methods.
Scope/intended purpose of models	To guide clinical decision-making and treatment planning.	

Note. PTSD = Post-traumatic stress disorder.

2.2.3 Information Sources, Searching, and Screening

Pre-defined search terms were used to search four databases: APA PsycInfo (via Ovid), PTSDpubs (via ProQuest), PubMed, and Scopus. The full search strategy is presented in Appendix B. No limits, restrictions, or filters were applied. Databases were searched on 27th April 2022. The following review articles were checked for potentially eligible studies: Aafjes van-Doorn et al. (2021), Chekroud et al. (2021), Chen et al. (2022), Dewar et al. (2020), Dwyer et al. (2018), Hahn et al. (2017), Glaz et al. (2021), Malgaroli and Schultebraucks (2021), Manchia et al. (2020), Meehan et al. (2022), Ramos-Lima et al. (2020). Forward and backward citation searches for all eligible studies were performed using *citationchaser* (Haddaway, 2021). The authors of all eligible studies were contacted to request further studies. Article metadata and abstracts for all search results were imported into EndNote 20 (https://endnote.com/). Duplicates automatically identified by EndNote 20 were screened and removed manually. Further duplicates were identified manually and removed during title and abstract screening. All titles and abstracts were manually screened against the inclusion and exclusion criteria in EndNote 20 by the first author, and full text files of potentially eligible studies were retrieved and screened.

2.2.4 Data Extraction and Synthesis

Relevant data from all eligible studies was extracted by the first author using a standardised data extraction table in Microsoft Excel, based on the Checklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS; Moons et al., 2014). This included sample characteristics, treatment details, methodological details, statistical outcomes, relevant findings, and authors' interpretation of findings. In some cases, study authors were contacted via email to clarify methods and results. Study characteristics, methods, and findings were tabulated and summarised using a narrative synthesis. The pre-registered intention was to quantitatively synthesise prediction model performance metrics using random effects meta-analysis, but this was not possible due to the heterogeneity of the study methods.

2.2.5 Risk of Bias Assessments

Risk of bias was assessed using the Prediction model study Risk Of Bias Assessment Tool (PROBAST; Moons et al., 2019). A second researcher independently conducted risk of bias assessments for 50% of the included studies, Cohen's kappa was calculated as a measure of agreement, discrepancies were discussed, and a third researcher was consulted where necessary.

2.3 Results

2.3.1 Study Selection

Detailed results of the study selection process are presented in the PRISMA diagram in Figure 2.1. In total, 1,570 titles and abstracts were screened, 48 potentially eligible full texts were screened, and 17 studies met the inclusion criteria for the review. Full texts that were screened and excluded are tabulated in Appendix C with reasons for exclusion.

2.3.2 Study Characteristics

Study characteristics are presented in Table 2.2. Most studies conducted a retrospective analysis of data (*k* = 12), either from clinical trials (*k* = 5), cohort studies (*k* = 1), or routine clinical practice (*k* = 6). Five studies prospectively collected data for analysis, either as a clinical trial (*k* = 1) or cohort study (*k* = 4). Five studies sampled any adults seeking treatment for PTSD; six sampled from military populations; five specified PTSD related to interpersonal-, childhood-, or sexual-abuse; and two sampled patients with co-occurring mental health problems (substance use disorder and depression, respectively). Participants received a range of CPG recommended psychological therapies for PTSD, most frequently PE (*k* = 10 studies), CPT (*k* = 6 studies), EMDR (*k* = 4 studies), or Tf-CBT (*k* = 3). Total sample size ranged from *N* = 57 to *N* = 612. All but one of the studies were published between 2018 and 2022. Nine studies took place in the USA, three in Germany, three in the Netherlands, one in Australia, and one was an analysis of data from England by a team of researchers in Germany and the UK.

Figure 2.1

PRISMA Flow Diagram

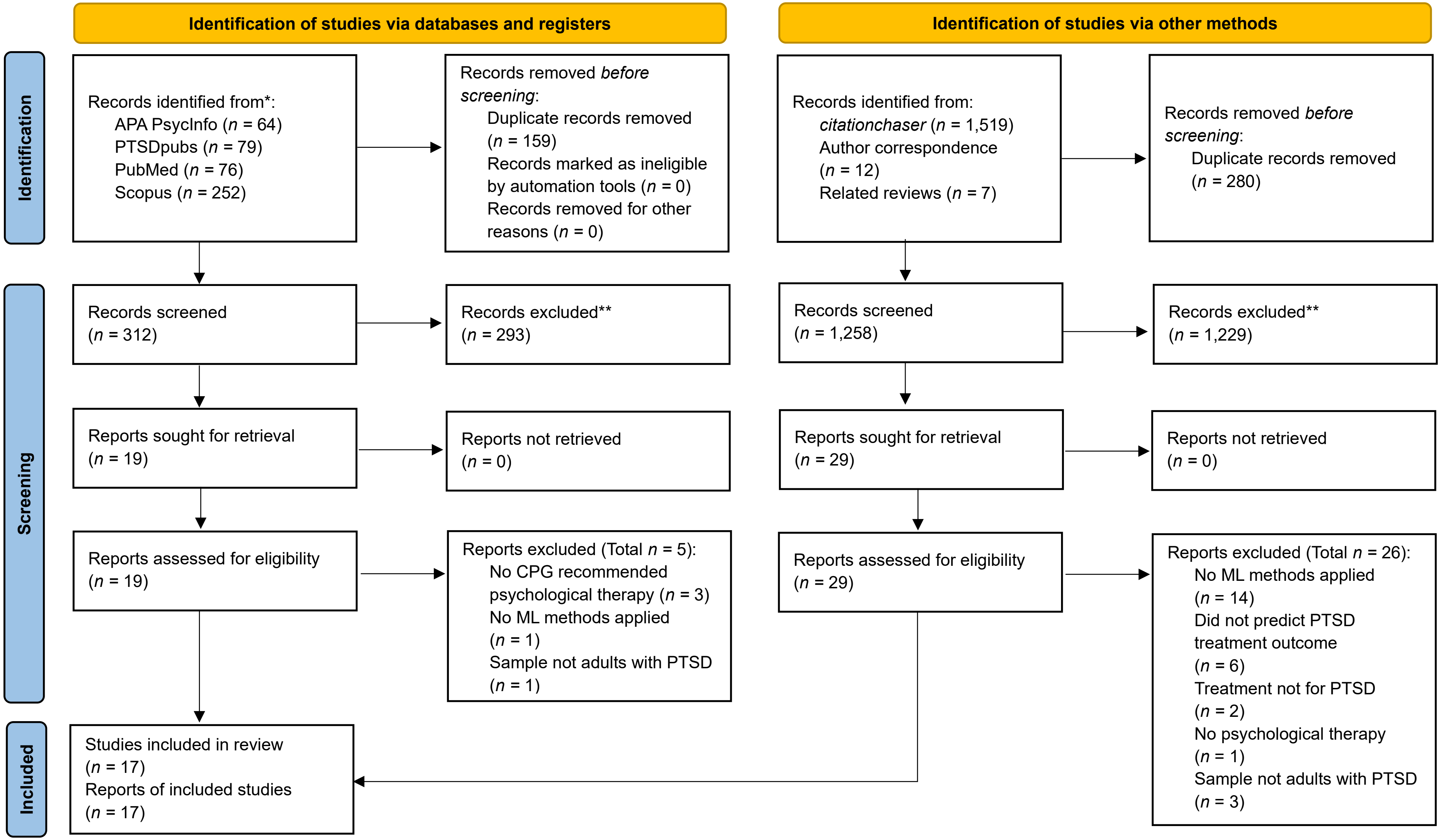

Identification of studies via databases and registers
Identification of studies via other methods
Identification
Screening
Included
Records identified from*:
APA PsycInfo (*n* = 64)
PTSDpubs (*n* = 79)
PubMed (*n* = 76)
Scopus (*n* = 252)
Records removed *before screening*:
Duplicate records removed (*n* = 159)
Records marked as ineligible by automation tools (*n* = 0)
Records removed for other reasons (*n* = 0)
Records identified from:
citationchaser (*n* = 1,519)
Author correspondence (*n* = 12)
Related reviews (*n* = 7)
Records removed *before screening*:
Duplicate records removed (*n* = 280)
Records screened (*n* = 312)
Records excluded** (*n* = 293)
Records screened (*n* = 1,258)
Records excluded** (*n* = 1,229)
Reports sought for retrieval (*n* = 19)
Reports not retrieved (*n* = 0)
Reports sought for retrieval (*n* = 29)
Reports not retrieved (*n* = 0)
Reports assessed for eligibility (*n* = 19)
Reports excluded (Total *n* = 5):
No CPG recommended psychological therapy (*n* = 3)
No ML methods applied (*n* = 1)
Sample not adults with PTSD (*n* = 1)
Reports assessed for eligibility (*n* = 29)
Reports excluded (Total *n* = 26):
No ML methods applied (*n* = 14)
Did not predict PTSD treatment outcome (*n* = 6)
Treatment not for PTSD (*n* = 2)
No psychological therapy (*n* = 1)
Sample not adults with PTSD (*n* = 3)
Studies included in review (*n* = 17)
Reports of included studies (*n* = 17)

Table 2.2

Study Characteristics

Study	Data Source	Population (Total sample *N*)	Setting (Country)	Treatment (Group *n*)	Treatment Duration
Deisenhofer et al. (2018)	Routine clinical practice (Retrospective)	Adults with PTSD (317)	NHS primary care outpatient mental health service (England)	Tf-CBT (242) EMDR (75)	≤ 20 weekly sessions (Session duration not reported)
Etkin et al. (2019)	RCT (Prospective)	Adults with PTSD (76)	University (U.S.A.)	PE (36) Wait-list control (30)	9 or 12 weekly or twice-weekly 90-minute sessions
Fleming et al. (2018)	Routine clinical practice (Retrospective)	Military veterans with PTSD (124)	Veterans Affairs speciality outpatient clinic (U.S.A.)	PE (49) CPT (53) Opted out of psychological therapy following introductory psychoeducation session (22)	Mean (SD) *n* sessions attended = 6.78 (7.03) (Session duration not reported)
Forbes et al. (2003)	Routine clinical practice (Retrospective)	Military veterans with PTSD (166)	Veterans PTSD treatment programme (Australia)	Group and individual therapy, primarily cognitive-behavioural in orientation, with trauma-focussed sessions (166)	16 sessions of individual therapy over 12 weeks (4 weeks inpatient, 8 weeks outpatient) (Session duration not reported)
Held et al. (2022)	Cohort study (Prospective)	Military veterans with PTSD (502)	University Medical Centre Intensive Outpatient Treatment Program (U.S.A.)	CPT based intensive PTSD treatment program (502)	14 once-daily 50-minute sessions of individual CPT over 3 weeks
Hendriks et al. (2018)	Cohort study (Prospective)	Adults with PTSD and history of multiple	Outpatient mental health clinic (Netherlands)	Intensive PE (73)	12 sessions over 4 days within 1 week (4.5 hours per-day), followed by 4 weekly 90-minute booster sessions with homework

Study	Data Source	Population (Total sample *N*)	Setting (Country)	Treatment (Group *n*)	Treatment Duration
		interpersonal traumas (73)			
Herzog et al. (2021)	Routine clinical practice (Retrospective)	Adults with PTSD (612)	Five specialised inpatient clinics (Germany)	Individual exposure therapy (PE, IRRT, or EMDR), plus group Tf-CBT and a range of supplementary psychological and non-psychological therapies (612)	8 to 10 weeks, 1 hour per week individual exposure therapy, 8 hours per week of group Tf-CBT, plus an average of 11 hours per week of multimodal and transdiagnostic interventions (total 152-200 therapy hours) Sample mean (SD, range) length of stay (days) = 54.3 (15.5, 6 - 98)
Hoeboer et al. (2021)	RCT (Retrospective)	Adults with childhood-abuse-related PTSD (149)	Two specialist outpatient mental health services (Netherlands)	PE (48) Intensified PE (51) STAIR+PE (50)	PE: 16 weekly 90-minute sessions Intensified PE: Three PE sessions per-week for 4 weeks, followed by booster PE sessions after 1 month and 2 months (total 14 sessions) STAIR+PE: Eight sessions of STAIR followed by eight sessions of PE
Keefe et al. (2018)	RCT (Retrospective)	Women with rape-trauma PTSD (160)	(U.S.A.)	CPT (79) PE (81)	Total 13 hours for each treatment over 6 weeks

Study	Data Source	Population (Total sample *N*)	Setting (Country)	Treatment (Group *n*)	Treatment Duration
					CPT: 12 sessions of 50-60 minutes, with 30 minutes added to each of the two writing exposure sessions (sessions 4 and 5) PE: Nine sessions; one 60-minute initial session followed by eight 90-minute sessions
Kratzer et al. (2019)	Routine clinical practice (Retrospective)	Inpatients with complex PTSD following childhood physical and childhood sexual abuse (150)	Specialist inpatient clinic (Germany)	Tf-CBT, often with integrated exposure and EMDR. Patients also offered group psychotherapies. (150)	≤ 20 individual psychological therapy sessions of 75-minutes each
López-Castro et al. (2021)	RCT (Retrospective)	Adults with PTSD and SUD (130)	Community based outpatient mental-health treatment programme (U.S.A.)	*Sample 1:* 1. COPE (33) 2. RPT (37) *Sample 2:* 1. Seeking Safety plus placebo (29) 2. Seeking Safety plus ADM (31)	*Sample 1:* All participants were offered 12 weekly 90-min individual sessions *Sample 2:* All participants were offered 12 weekly 60-min individual psychological therapy sessions, and ADM (sertraline) dosage started on 50 mg/day and increased

Study	Data Source	Population (Total sample *N*)	Setting (Country)	Treatment (Group *n*)	Treatment Duration
					up to 200 mg/day over 2 weeks throughout the active study period
Nixon et al. (2021)	RCT (Retrospective)	Female interpersonal trauma survivors (216)	Community (U.S.A.)	CPT (216)	12 weekly or bi-weekly 60-min sessions
Stirman et al. (2021)	RCT (Retrospective)	Female military veterans and active-duty service members with PTSD (267)	Nine VA medical centres, two VA readjustment counselling centres, and a military hospital (U.S.A.)	PE (135) Present-Centred Therapy (132)	10 weekly 90-minute sessions
Stuke et al. (2021)	Routine clinical practice (Retrospective)	Adults with PTSD (209)	Specialist day clinic (Germany)	CBT based day-care programme including individual CPT (209)	Four sessions per-week of individual CPT, plus group trauma-focussed therapy 5 days per-week, for a mean of 8.59 weeks (SD = 1.4) (Session duration not reported)
Zhang et al. (2021)	Cohort study (Prospective); non-randomised clinical trial (Prospective)	Military veterans with PTSD (241); trauma-exposed controls (95)	University; Veterans Affairs PSTD clinic (U.S.A.)	PE or CPT (135)	Based on published, manualised protocols (Number of sessions and session duration not reported)

Study	Data Source	Population (Total sample *N*)	Setting (Country)	Treatment (Group *n*)	Treatment Duration
Zhutovsky et al. (2019)	Cohort study (Prospective)	Male military veterans with PTSD (57); combat-exposed controls (29)	Four military mental-healthcare outpatient clinics (Netherlands)	Tf-CBT (8) EMDR (28) Tf-CBT+EMDR (8)	Mean (SD) number of treatment sessions: Responders = 9.86 (6.29) Non-responders = 10.05 (4.22) (Session duration not reported)
Zilcha-Mano et al. (2020)	Cohort study (Retrospective)	Adults with PTSD (51); adults with PTSD and depression (52); trauma-exposed controls (76)	State Psychiatric Institute (U.S.A.)	PE (55)	10-week standard PE protocol (Session duration not reported)

Note. ADM = anti-depressant medication; CBT = cognitive behavioural therapy; COPE = concurrent treatment for substance use disorder and post-traumatic stress disorder combining prolonged exposure and relapse prevention therapy; CPT = cognitive processing therapy; EMDR = eye movement desensitization and reprocessing; IRRT = imagery rescripting and reprocessing therapy; NHS = National Health Service; PE = prolonged exposure; PTSD = post-traumatic stress disorder; RCT = randomised control trial; RPT = relapse prevention therapy (treatment for substance use disorder); Seeking Safety = skills-based intervention for concurrent post-traumatic stress disorder and substance use disorder; STAIR = Skills Training in Affective and Interpersonal Regulation; SUD = substance use disorder; Tf-CBT = Trauma-focussed cognitive behavioural therapy; VA = Veterans Affairs.

2.3.3 Risk of Bias Assessments with PROBAST

Detailed risk of bias assessments are tabulated in Appendix D. The first and second rater initially agreed on seven out of nine studies, corresponding to a Cohen's kappa = 0.4, indicating *fair* agreement. Following consultation with a third researcher consensus was reached on all nine studies. All seventeen studies were rated at high risk of bias overall, primarily as all studies were high risk of bias in the *Analysis* domain. None of the studies had an adequate number of participants with the outcome, and for some studies the number of predictor parameters estimated was unclear (studies often reported the number of candidate variables but did not report dummy coding of categorical variables or whether psychometric measures were entered as total, factor-level, or item-level scores). Although nine studies reported metrics of prediction accuracy, error, and/or discrimination, none of the studies reported calibration and therefore relevant model performance metrics were not evaluated appropriately. Thirteen studies did not include all enrolled participants in the analysis. Three studies inappropriately handled missing data and six studies did not provide information on the handling of missing data. Seven studies were rated at risk of bias due to selection of participants for using routinely collected clinical data or retrospective cohort study data.

2.3.4 Study Methods and Results

Study methods are tabulated in Table 2.3 and results are tabulated in Appendix E.

Table 2.3

Study Methods

Study	Outcome(s) to be predicted (variable type, measure)	Outcome Time-Point	Predictor type (number of candidate predictor variables)	Machine learning methods (purpose, *n* analysed, *n* with outcome)	Additional methods (purpose, *n* analysed)	Sample size calculation	Data pre-processing	Hyper-parameter setting	Validation methods	Evidence level
Deisenhofer et al. (2018)	Post-treatment symptom severity (continuous, PHQ-9 as a proxy measure of PTSD) Optimal treatment for each patient (PAI)	Final treatment session	Clinical, demographic, psychometric (11)	Genetic algorithm (feature selection, *n* = 150; 75)	Linear regression (parameter estimation, calculate PAI, *n* = 150; 75) Chi-squared test (compare rate of reliable improvement between PAI indicated optimal vs. suboptimal treatment, *n* = 225)	NR	Multiple imputation via random forest (on whole sample) Categorical predictors reduced to dichotomous variables (employment, medication) Propensity score matching	Genetic algorithm variable importance threshold set at 80% Other hyper-parameter settings not reported	Leave-one-out cross-validation	2
Etkin et al. (2019)	≥50% reduction in PTSD score (binary, CAPS)	4 weeks after final treatment session	MRI, EEG, neurocognitive tests (number of candidate predictor variables unclear)	Linear support vector machine; Non-linear radial basis function support vector machine (predict	Generalised linear modelling (neurocognitive feature selection, *n* = 92 including *n* = 36 controls;	NR	Threshold in delayed recall score indicative of impaired recall	NR	Leave-one-out cross-validation	2

Study	Outcome(s) to be predicted (variable type, measure)	Outcome Time-Point	Predictor type (number of candidate predictor variables)	Machine learning methods (purpose, *n* analysed, *n* with outcome)	Additional methods (purpose, *n* analysed)	Sample size calculation	Data pre-processing	Hyper-parameter setting	Validation methods	Evidence level
				treatment outcome, *n* = 36, number of participants with the outcome not reported)	neuroimaging feature selection, *n* = 87 including *n* = 36 healthy controls) Generalised linear mixed modelling (test interactions with treatment, *n* = 36, vs. control, *n* = 30)		identified by discriminant analysis (*n* = 92) Preprocessing of neuroimaging data described in supplement			
Fleming et al. (2018)	Retention (count, *n* sessions completed)	Final treatment session	Clinical, demographic, psychometric, military service characteristics, trauma characteristics (51)	Exhaustive CHAID classification tree (feature selection, parameter estimation, prediction, *n* = 122)		NR	NR	NR	NR	1
Forbes et al. (2003)	Change in symptom score (continuous, PCL)	3 months post-treatment; 9 months post-treatment (*n* = 136)	Psychometric (16)	*k*-means cluster analysis (test reliability of subgroups identified by Ward's cluster analysis, *n* = 158)	Ward's hierarchical cluster analysis (identify subgroups of PTSD patients, *n* =158)	NR	NR	NR	NR	1

Study	Outcome(s) to be predicted (variable type, measure)	Outcome Time-Point	Predictor type (number of candidate predictor variables)	Machine learning methods (purpose, *n* analysed, *n* with outcome)	Additional methods (purpose, *n* analysed)	Sample size calculation	Data pre-processing	Hyper-parameter setting	Validation methods	Evidence level
					Second order principal components analysis (reduce MMPI-2 scale and aid interpretation of results, *n* = 158) Multivariate generalised linear modelling (explore differences in outcome and independent variables between clusters, *n* = 158) Repeated measures multivariate generalised linear modelling (examine differences in treatment response between subgroups, *n* = 158)					

Study	Outcome(s) to be predicted (variable type, measure)	Outcome Time-Point	Predictor type (number of candidate predictor variables)	Machine learning methods (purpose, *n* analysed, *n* with outcome)	Additional methods (purpose, *n* analysed)	Sample size calculation	Data pre-processing	Hyper-parameter setting	Validation methods	Evidence level

Held et al. (2022)	Minimal response (binary, PCL-5);
	Fast response (binary, PCL-5)
		Intake, treatment days 2, 3, 5, 6, 8, 11, and 13, and post-treatment	Demographic, psychometric, military service characteristics, trauma characteristics (104)	Elastic Net classification; Gradient Boosted Models; Random Forest;
				Ridge classification;
				Logistic Regression with Max-Min Parent-Child variable selection (feature selection, parameter estimation, prediction, *n* = 432 including *n* = 73 with minimal response outcome and *n* = 61 with fast response outcome)
					Group-based trajectory modelling (identify response trajectory class)
					Logistic Regression (comparison with ML methods)
						NR	Listwise exclusion of participants with missing data
							One-hot-encoding of categorical variables
							Performance assessed by area under the precision-recall curve to account for class imbalance
								Optimisation via five-fold cross-validated grid search within inner loop of nested five-fold cross validation
								Hyper-parameter tuning not required for logistic regression or logistic regression with max-min parent-child variable selection
									Five-fold cross-validation	2

Study	Outcome(s) to be predicted (variable type, measure)	Outcome Time-Point	Predictor type (number of candidate predictor variables)	Machine learning methods (purpose, *n* analysed, *n* with outcome)	Additional methods (purpose, *n* analysed)	Sample size calculation	Data pre-processing	Hyper-parameter setting	Validation methods	Evidence level
Hendriks et al. (2018)	Response trajectory class (polytomous, CAPS)	Baseline, 3-month follow up, 6-month follow up	Clinical, demographic, psychometric (14)	*k*-means cluster analysis (identify response trajectory class, *n* = 69)	Stepwise multinominal logistic regression (feature selection and prediction, *n* = 69)	NR	Multiple imputation of missing data following a framework for multiple imputation in cluster analysis Participants missing baseline CAPS score were excluded (*n* = 4)	Varied number of clusters from 3 to 6 and evaluated goodness of fit Other hyper-parameter settings not reported	NR	1
Herzog et al. (2021)	Change in symptom score (continuous, IES-R)	First and last day of treatment	Clinical, demographic, psychometric (≥46)	Elastic net (feature selection, parameter estimation, prediction, *n* = 397)		NR	Participants missing >60% and variables missing >40% were excluded Univariate outlier values removed	L1 and L2 penalty weighting alpha set to 0.5 Optimal lambda value estimated by	Tested on randomly partitioned (35%) hold-out validation set (*n* = 215) Bootstrap internal cross-	3

Study	Outcome(s) to be predicted (variable type, measure)	Outcome Time-Point	Predictor type (number of candidate predictor variables)	Machine learning methods (purpose, *n* analysed, *n* with outcome)	Additional methods (purpose, *n* analysed)	Sample size calculation	Data pre-processing	Hyper-parameter setting	Validation methods	Evidence level
							Time-event data log-transformed Categorical variables were reduced to binary or continuous variables (details not reported), ICD-10 medical diagnoses were dummy coded Binary variables with class imbalance were excluded Multiple imputation via	*k*-fold cross-validation averaged across 10 repetitions (within training set)	validation in training set (*n* = 397)	

Study	Outcome(s) to be predicted (variable type, measure)	Outcome Time-Point	Predictor type (number of candidate predictor variables)	Machine learning methods (purpose, *n* analysed, *n* with outcome)	Additional methods (purpose, *n* analysed)	Sample size calculation	Data pre-processing	Hyper-parameter setting	Validation methods	Evidence level
							random forest (separately on training and test set)			
Hoeboer et al. (2021)	Change in symptom score (continuous, CAPS-5; PCL-5) Optimal treatment for each patient (PAI)	4 weeks, 8 weeks, and 16 weeks after start of treatment	Clinical, demographic, psychometric (24)	Boruta algorithm random forest classifier (feature selection, *n* = 99; 50)	Linear mixed-effect modelling (estimate change in symptoms over the course of treatment for each participant, *n* = 149) Linear regression (parameter estimation, prediction, *n* = 99; 50)	NR	NR	NR	Bootstrapping (feature selection) Leave-one-out cross-validation internal cross-validation (prediction, PAI)	2
Keefe et al. (2018)	Dropout (binary, treatment completion) Optimal treatment for each patient (PAI)	Final treatment session	Clinical, demographic, psychometric, trauma characteristics (20)	Bootstrapped, random forest variant of model-based recursive partitioning, and bootstrapped variant of an AIC-based backward selection model	Logistic regression (parameter estimation, prediction, *n* = 160)	NR	Participants who dropped out prior to randomisation excluded from analyses (*n* = 11)	NR	Five-fold cross-validation	2

Study	Outcome(s) to be predicted (variable type, measure)	Outcome Time-Point	Predictor type (number of candidate predictor variables)	Machine learning methods (purpose, *n* analysed, *n* with outcome)	Additional methods (purpose, *n* analysed)	Sample size calculation	Data pre-processing	Hyper-parameter setting	Validation methods	Evidence level
				(feature selection, *n* = 160 including *n* = 49 with dropout outcome)			Single-dataset random forest imputation strategy using all available pre-treatment and outcome data			
Kratzer et al. (2019)	Reliable change (binary, IES-R)	Before discharge	Clinical, psychometric (5)	Conditional inference tree (feature selection and prediction, *n* = 150 including *n* = 78 with reliable change outcome)		NR	Bayesian multiple imputation	NR	NR	1
López-Castro et al. (2021)	Treatment attendance (count, *n* sessions attended)	Final treatment session	Clinical, demographic, psychometric, trauma characteristics (28)	Iterative Random Forest (feature selection, *n* = 70)	Poisson regression (parameter estimation, prediction, *n* = 70; 60)	NR	NR	Default hyper-parameter settings used	Parameter estimation repeated in second dataset (*n* = 70; *n* = 60)	1
Nixon et al. (2021)	Response trajectory class (polytomous, PDS/PSS)	Post-treatment, follow up 3 to 9 months	Clinical, demographic, psychometric, trauma characteristics (38)	Random forests of conditional inference trees (feature selection		NR	Response trajectories identified based on symptom	Default hyper-parameter settings used	Random forest (bagging)	2

Study	Outcome(s) to be predicted (variable type, measure)	Outcome Time-Point	Predictor type (number of candidate predictor variables)	Machine learning methods (purpose, *n* analysed, *n* with outcome)	Additional methods (purpose, *n* analysed)	Sample size calculation	Data pre-processing	Hyper-parameter setting	Validation methods	Evidence level
		after final session		and prediction, *n* = 179)			scores at session 1, session 6, posttreatment and follow-up			
Stirman et al. (2021)	Post-treatment symptom severity (continuous, CAPS) Optimal treatment for each patient (PI)	Post-treatment	Clinical, demographic, psychometric, trauma characteristics (29)	Elastic net, five iterations, features retained if selected on all five iterations. Then stepwise AIC-penalised bootstrapped variable selection with 10,000 bootstrapped samples, features retained if selected in >60% samples (feature selection, *n* = 267)	Linear regression with 10-fold cross-validation, coefficients mean averaged across 1000 runs (parameter estimation, generate PI, *n* = 267) Linear regression (test association between PI and outcome, and interaction between PI and treatment type, *n* = 267)	NR	Binary variables effect-coded Continuous predictors standardised Multiple imputation via random forest (OOB error estimates reported)	Elastic net alpha parameter set to .75, lambda optimised via 10-fold cross-validation	10-fold cross-validation	2
Stuke et al. (2021)	Change in symptom score	Discharge	Clinical, demographic, psychometric,	Principal component analysis (feature reduction, *n* = 115)	Linear regression (comparison with ADAboost regressor, *n* = 115)	NR	Participants missing responses to a whole scale	Optimal number of components for each	Leave-one-out cross-validation	2

Study	Outcome(s) to be predicted (variable type, measure)	Outcome Time-Point	Predictor type (number of candidate predictor variables)	Machine learning methods (purpose, *n* analysed, *n* with outcome)	Additional methods (purpose, *n* analysed)	Sample size calculation	Data pre-processing	Hyper-parameter setting	Validation methods	Evidence level
	(continuous, DTS)		trauma characteristics (12)	ADAboost regressor (parameter estimation, prediction, *n* = 115)			excluded; scale mean imputed where participants were missing <20% responses to scale (*n* = 10)	participant estimated via hyper-parameter optimisation with 10-fold cross-validation in (*N* - 1) training set, varying number of components from 1-10 and comparing squared error ADAboost: *n* estimators optimised with 10-fold cross-validation in training set		

Study	Outcome(s) to be predicted (variable type, measure)	Outcome Time-Point	Predictor type (number of candidate predictor variables)	Machine learning methods (purpose, *n* analysed, *n* with outcome)	Additional methods (purpose, *n* analysed)	Sample size calculation	Data pre-processing	Hyper-parameter setting	Validation methods	Evidence level
								(candidates: 2, 5, 10, 20, 40); default settings used for other hyper-parameters		
Zhang et al. (2021)	Post-treatment symptom severity (continuous, CAPS; CAPS-5)	NR for PTSD data	EEG/PEC (unclear)	Sparse *k*-means clustering (identify PTSD subtypes, *n* = 106)	Linear mixed models (predict outcome from subtype, *n* = 72; *n* = 63)	NR	Multiple imputation reported for depression dataset but not for PTSD dataset EEG and MRI preprocessing	Number of clusters determined and assessed by statistical criteria (the gap statistic) Sparsity parameter determined by inner-loop cross-validation	*k*-means repeated on 100 randomly selected (90%) subsample PTSD treatment outcomes dataset divided into two cohorts, cluster analysis applied, and linear mixed modelling repeated in the second cohort	1

Study	Outcome(s) to be predicted (variable type, measure)	Outcome Time-Point	Predictor type (number of candidate predictor variables)	Machine learning methods (purpose, *n* analysed, *n* with outcome)	Additional methods (purpose, *n* analysed)	Sample size calculation	Data pre-processing	Hyper-parameter setting	Validation methods	Evidence level
Zhutovsky et al. (2019)	≥30% reduction in PTSD score (binary, CAPS)	6 to 8 months from baseline assessment	MRI (unclear)	Independent component analysis using the meta-ICA approach (dimension reduction, *n* = 28 controls) Gaussian process classifier (feature selection and prediction, *n* = 44 including *n* = 20 with treatment response outcome)	Univariate analysis with threshold-free cluster enhancement and permutation analysis (dimension reduction, *n* = 44)	NR	Participants missing follow-up data were excluded from analysis, and 3 participants were excluded due to excessive movement during MRI MRI pre-processing reported in supplement	NR	10-fold cross-validation	2
Zilcha-Mano et al. (2020)	Change in symptom score (continuous, CAPS)	Pre to post-treatment	MRI (unclear)	Linear kernel support vector machine with *t*-test filtering and wrapper based sequential feature selection (feature	Pearson correlations (test correlation between features and treatment outcome, *n* = 55)	NR	Excluded 3 participants due to excessive movement during MRI Features	Hyper-parameter optimisation (kernel scale and function) during 10-	10-fold cross-validation during support vector machine training Correlations not cross-validated	1

Study	Outcome(s) to be predicted (variable type, measure)	Outcome Time-Point	Predictor type (number of candidate predictor variables)	Machine learning methods (purpose, *n* analysed, *n* with outcome)	Additional methods (purpose, *n* analysed)	Sample size calculation	Data pre-processing	Hyper-parameter setting	Validation methods	Evidence level
				reduction and selection, *n* = 179)			regressed for age, sex, and MRI scanner, and normalised MRI pre-processing reported in supplement	fold cross-validation		

Note. AIC = Akaike Information Criterion; CAPS = Clinician-Administered PTSD Scale; CAPS-5 = Clinician-Administered PTSD Scale for DSM-5; CHAID = Chi-square Automatic Interaction Detection; DTS = Davidson Trauma Scale; EEG = Electroencephalography; ICD-10 = International Classification of Diseases 10th Revision; IES-R = Impact-of-Event-Scale-Revised; MMPI-2 = Minnesota Multiphasic Personality Inventory-2; MRI = Magnetic Resonance Imaging; NR = Not Reported; OOB = Out-Of-Bag; PAI = Personalised Advantage Index; PCL = PTSD Checklist; PCL-5 = PTSD Checklist for DSM-5; PDS = Posttraumatic Stress Diagnostic Scale; PEC = Power Envelope Connectivity; PHQ-9 = Patient Health Questionnaire-9; PI = Prognostic Index; PSS = Post-traumatic Symptoms Scale; PTSD = Post-traumatic Stress Syndrome; RCT = Randomised Controlled Trial.

2.3.4.1 Outcome Variable

Fourteen studies sought to predict treatment response, operationalised in a variety of different ways. Eight studies sought to predict treatment response as a continuous outcome, five of which predicted change in PTSD score from pre- to post-treatment, two predicted post-treatment PTSD score, and one predicted post-treatment depression score as a proxy outcome (Deisenhofer et al., 2018). Six studies sought to predict treatment response as a categorical outcome, two of which predicted percentage change in PTSD score (50% and 30% respectively) as a binary outcome, one predicted reliable change in PTSD score as a binary outcome, two predicted latent trajectory class membership as a polytomous outcome (Hendriks et al., 2018; Nixon et al., 2021), and one predicted latent trajectory class membership as two binary outcomes (Held et al., 2022). The remaining three studies sought to predict treatment retention, two of which predicted a count of the number of sessions attended (Fleming et al., 2018; López-Castro et al., 2021), and one predicted dropout as a binary outcome (Keefe et al., 2018).

2.3.4.2 Candidate Predictor Variables

Thirteen studies employed psychometric data (e.g., self-report or clinician-report measures of PTSD, depression, anxiety) as candidate predictor variables, eleven of these also used demographic data (e.g., gender, age, employment status), and eleven also used clinical data (e.g., diagnoses, medication use). Eleven studies tested baseline PTSD symptoms and PTSD related cognitions as candidate predictors, and seven of these also tested trauma characteristics such as type of trauma and time since trauma. Four studies explored the relationship between neuroimaging data (MRI and EEG) and PTSD treatment outcomes. Number of candidate predictor variables ranged from approximately 5 to 104. Studies that used neuroimaging data did not specify the number of candidate predictors. See Appendix F for details of candidate predictor variables.

2.3.4.3 Predictors Included in the Final Model

Among the fourteen studies that sought to predict treatment response, all but one (Nixon et al., 2021) reported at least one significant pre-treatment predictor. Five studies included PTSD severity as a predictor in the final model, two of which found that specific PTSD symptoms or symptom clusters were important predictors (Held et al., 2022; Herzog et al., 2021). Four of these also included trauma related variables such as type of trauma (Hoeboer et al., 2021; Stirman et al., 2021), post-traumatic cognitions (Held et al., 2021; Stuke et al., 2021), and perceived centrality of trauma to person's identity (Stuke et al., 2021). Six studies included co-occurring mental health problems such as depression (*k* = 5) and emotion regulation difficulties (*k* = 2); and five included demographic variables such as age (*k* = 3) and gender (*k* = 3). Three studies using MRI data identified regions of the brain associated with treatment response, but there was no consensus between them (Etkin et al., 2019; Zhutovsky et al., 2019; Zilcha-Mano et al., 2020). Studies found that PTSD, trauma, and mental health related variables were stronger predictors of treatment response than demographic variables (Held et al., 2022; Herzog et al., 2021; Hoeboer, Oprel, et al., 2021; Stirman et al., 2021; Stuke et al., 2021). There was little consensus among the predictors included in the final model for the three studies that sought to predict treatment retention or dropout (Fleming et al., 2018; Keefe et al., 2018; López-Castro et al., 2021), but all found that trauma related variables were important predictors, specifically experiences of abuse (Keefe et al., 2018), time since trauma (López-Castro et al., 2021), and motivational readiness to address trauma (Fleming et al., 2018).

2.3.4.4 Machine Learning Methods

Studies used a range of different ML methods for various purposes. Fourteen studies used supervised ML methods. Eight studies used decision tree-based methods, and all but two of these used ensemble tree methods such as random forest and boosting algorithms (ADAboost, gradient boosted models). Three studies used a penalised regression method called *elastic net* (Held et al., 2022; Herzog et al., 2021; Stirman et al., 2021). Three studies used kernel methods (support vector machine, Gaussian process classifier) to analyse MRI data (Etkin et al., 2019; Zhutovsky et al., 2019; Zilcha-Mano et al., 2020). Five studies used unsupervised clustering (*k*-means) or dimension reduction methods (principal component analysis, independent component analysis). None of the studies used Bayesian ML methods or deep learning methods such as neural networks.

Five studies used the same ML method to perform feature selection, parameter estimation, and outcome prediction (Fleming et al., 2018; Held et al., 2022; Herzog et al., 2021; Kratzer et al., 2019; Nixon et al., 2021). Two studies used an unsupervised ML method for feature reduction and then used a supervised ML method for prediction (Stuke et al., 2021; Zhutovsky et al., 2019). Five studies used supervised ML methods to select predictors, and then used simpler statistical methods (e.g., linear regression, correlation) to test the relationship between the selected predictors and outcome (Hoeboer, Oprel, et al., 2021; Keefe et al., 2018; López-Castro et al., 2021; Stirman et al., 2021; Zilcha-Mano et al., 2020). One study used a genetic algorithm to select predictors for a linear regression model (Deisenhofer et al., 2018). One study used generalised linear modelling to select predictors and then used supervised ML methods to predict outcomes (Etkin et al., 2019).

Three studies used *k*-means cluster analysis. Zhang et al. (2021) used *k*-means to identify PTSD subtypes and then linear mixed models to test the relationship between subtypes and treatment outcome. Hendriks et al. (2018) used *k*-means to identify treatment response trajectory classes, and then used stepwise logistic regression to select predictors and predict trajectory class membership. Forbes et al. (2003) used *k*-means to test the reliability of PTSD subtypes identified by Ward's hierarchical cluster analysis, and then used generalised linear modelling to test differences in treatment response between subtypes.

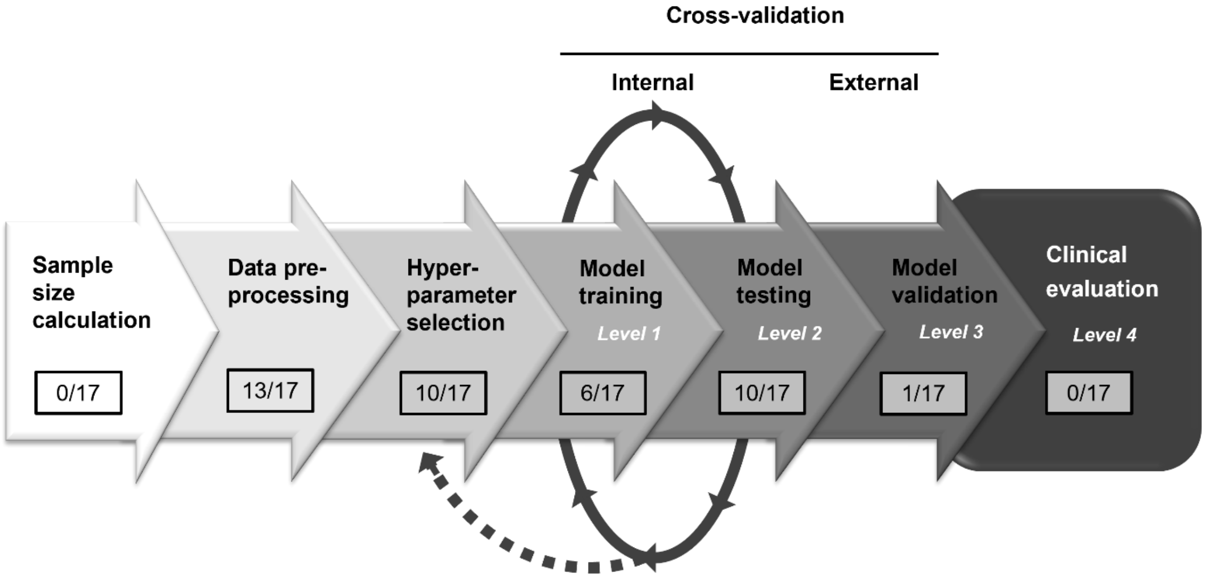
Two studies compared the performance of more than one ML method (Etkin et al., 2019; Held et al., 2022), and two studies compared the performance of ML methods against that of traditional statistical methods (Held et al., 2022; Stuke et al., 2021).

2.3.4.5 Adherence to the Machine Learning Pipeline

The number of studies that reported each step of the ML pipeline is presented in Figure 2.2.

Figure 2.2

Proportion of Studies that Reported Each Step of the Machine Learning Pipeline



Note. Figure adapted from Delgadillo and Atzil-Slonim (2022)

2.3.4.5.1. Sample Size Calculation

None of the studies reported a sample size calculation. The number of participants with the outcome in a training sample ranged from < 36 (Etkin et al., 2019) to 397 (Herzog et al., 2021).

2.3.4.5.2. Data Pre-processing

Nine studies reported handling of missing data, six of which reported multiple imputation. Three studies performed multiple imputation via random forest, but only one reported out-of-bag error estimates (Stirman et al., 2021). One study reported listwise exclusion of participants with missing data (Held et al., 2022); one excluded participants missing follow-up data (Zhutovsky et al., 2019); one excluded participants missing a whole scale and imputed mean values where <20% of a scale was missing (Stuke et al., 2021). Three studies reported reduction of categorical variables, one reported transformation of variables, one reported handling of class imbalance, and one reported case-control matching. Three of four studies that used neuroimaging data reported preprocessing of neuroimaging data. Four studies did not report any pre-processing of data.

2.3.4.5.3 Hyperparameter Selection

Six studies reported using internal-cross validation to optimise hyperparameter settings, one of which also reported using default settings for some hyperparameters (Stuke et al., 2021). Two studies only reported using default hyperparameter settings (López-Castro et al., 2021; Nixon et al., 2021). Two studies reported using statistical criteria (goodness of fit, gap statistic) to decide the number of k-means clusters (Hendriks et al., 2018; Y. Zhang et al., 2021). Some studies reported selection for some but not all hyperparameters, and seven studies did not report hyperparameter selection.

2.3.4.5.4 Cross-validation and Level of Evidence

Ten studies internally cross-validated predictions: four performed *k*-fold, four performed leave-one-out, and two performed bootstrapping. One study also performed external validation in a randomly partitioned hold-out dataset (Herzog et al., 2021). As such, ten studies provided level 2 evidence and one study provided level 3 evidence.

Six studies did not internally or externally cross-validate model predictions and therefore provided level 1 evidence. One of these studies performed k-fold during predictor selection but not during prediction (Zilcha-Mano et al., 2020). One study used the predictors selected in one dataset to make predictions in a second dataset, but repeated parameter estimation (model fitting) in the second dataset, and therefore performed replication rather than external validation (López-Castro et al., 2021). Another study divided the dataset into two cohorts and repeated *k*-means clustering and linear mixed modelling in the second cohort, again performing replication rather than external validation (Y. Zhang et al., 2021).

2.3.4.6 Evaluation Metrics

Nine studies reported metrics of model prediction accuracy or error. These studies all applied internal cross-validation procedures, but it is important to note that only Herzog et al. (2021) performed external validation, and none had a reasonable number of participants with the outcome. Therefore, model performance metrics were estimated within a training sample of insufficient size, limiting the likelihood that they will generalise to independent samples. None of the studies that sought to predict treatment retention reported evaluation metrics. None of the studies reported calibration.

Among the eight studies that sought to predict a continuous outcome, three reported model prediction accuracy/discrimination in the form of R^2 or R, and four reported prediction error in the form of root mean squared error (RMSE) or mean absolute error (MAE). Two of these studies reported both accuracy and error, and four studies did not report either. Herzog et al. (2021) used elastic net and reported $R^2 = 0.17$ (MAE = 0.69, RMSE = 0.91) in the training set (with bootstrap internal-cross validation) and $R^2 = 0.16$ (MAE = 0.77, RMSE = 0.95) in the hold-out external validation. Stirman et al. (2021) used elastic net to select predictors and reported $R^2 = 0.39$ (RMSE = 20.28) for prediction with linear regression mean averaged over 1000 repetitions of 10-fold internal cross-validation. Stuke et al. (2021) used principal component analysis to select predictors and reported $R = 0.162$ for prediction with ADAboost regressor and $R = 0.214$ for linear regression (when squared, ADAboost $R^2 = 0.03$ and linear regression $R^2 = 0.05$). Hoeboer et al. (2021) reported RMSE ranging from 4.06 to 7.24 when predicting change on two PTSD measures in two treatment groups (RMSE is referred to as *average error* in the publication and was clarified through correspondence with the author). Deisenhofer et al. (2018) reported *true error* (MAE of factual predictions) of 4.92 in one treatment group and 5.37 in the other.

Among the six studies that sought to predict a categorical outcome, two reported accuracy as raw accuracy or balanced accuracy, and three reported discrimination as area under the receiver operating characteristic curve (AUC-ROC), area under the precision-recall curve (AUC-PR), and/or sensitivity and specificity. Nixon et al. (2021) visually examined AUC-ROC but did not report statistics, and a further two studies did not report evaluation metrics for prediction of categorical outcomes. Held et al. (2022) tested six methods of developing a classification model and found that gradient boosted models produced the best predictions of fast response (AUC-PR = 0.466, AUC-ROC = 0.765) and elastic net produced the best predictions of minimal response (AUC-PR = 0.628, AUC-ROC = 0.826). Zhutovsky et al. (2019) used Gaussian process classifier to predict ≥ 30% reduction in PTSD score from MRI data and reported AUC-ROC = 0.929, balanced accuracy = 81.4%, sensitivity = 84.8%, specificity = 78%. Etkin et al. (2019) predicted ≥ 50% reduction in PTSD score from verbal memory delayed recall impairment and low within Ventral Attention Network connectivity

(MRI) and reported accuracy = 85%, sensitivity = 80%, and specificity = 87% for linear SVM, and accuracy = 90%, sensitivity = 80%, and specificity = 93% for radial basis function SVM, but the sample size was particularly small (*n* = 36), the number of participants with the outcome was not reported, and class imbalance was not addressed.

2.3.4.7 Predicting Differential Treatment Outcome

Five studies explored interactions between pre-treatment variables and treatment type. Three studies sought to retrospectively predict the optimal treatment for each participant by developing a personalised advantage index (Deisenhofer et al., 2018; Hoeboer et al., 2021; Keefe et al., 2018). Following a method suggested by Kessler et al. (2017), Deisenhofer et al. (2018) and Hoeboer et al. (2021) used ML methods to select predictors for a linear regression model for each treatment under investigation and identified each patients' optimal treatment by comparing the outcomes predicted by the two regression models. Both studies found a significantly greater improvement in symptoms among patients who had received their model indicated optimal treatment. Keefe et al. (2018) used ML methods to select predictors and moderators (i.e., variables that interact with treatment type) for a logistic regression model and found a significantly lower rate of dropout among patients who received their model-indicated optimal treatment.

Stirman et al. (2021) sought to identify patients most likely to benefit from the most efficacious of two treatments, and those for whom treatment type was unlikely to make a difference, by developing a prognostic index (composite predictor) and testing the interaction between the prognostic index and treatment type. The interaction explained 39% of the variance in post-treatment PTSD severity. All four of the above studies reported that using ML methods in this way could potentially guide personalised treatment selection for PTSD.

Zhang et al. (2021) investigated whether patients with latent subtypes of PTSD identified via *k*-means of EEG data, and not identifiable through clinical measures or demographic data, responded differentially to two treatments. There was a significant difference in post-treatment severity between the two subtypes, but no interaction with treatment type. Patients in this study were not randomly allocated to treatment and this was not addressed, therefore there is potential confounding by indication (Kyriacou & Lewis, 2016).

2.4 Discussion

This review systematically reviewed studies that used ML methods to predict the outcome of psychological therapies for PTSD, and the degree to which studies adhered to the best practice ML pipeline. Through searching four databases and eleven similar systematic reviews, conducting forward and backward citation searches, and contacting the authors of eligible papers, seventeen studies were identified that met the inclusion criteria. Sixteen were published within the previous four years, reflecting a recent surge of interest in ML methods in clinical psychology and psychiatry (Chekroud et al., 2021). The one exception was published almost 20 years earlier, but this study made no reference to ML and simply used *k*-means to test the reliability of clusters identified via a different clustering method (Forbes et al., 2003). Risk of bias assessments using PROBAST found all studies to be at high risk of bias, notably due to inadequate sample size. The number of participants in a training sample ranged from < 36 to 397. Studies applied a diverse range of ML methods. Fourteen studies used supervised ML methods, eight of which used decision tree-based methods, and six of these used ensemble tree methods such as random forest. Five studies used unsupervised methods, three of which used *k*-means. Regarding the ML pipeline, none of the studies reported a sample size calculation, seven studies did not report hyperparameter setting, six did not report internal cross-validation, and only one study performed external validation.

2.4.1 Considerations Regarding Risk of Bias

All studies were rated high risk of bias in the *analysis* domain, primarily due to inadequate sample size and neglecting to assess model calibration. Six studies were rated high risk of bias in the *participants* domain for using routinely collected practice data. Moons et al. (2019) suggest that routinely collected data is at higher risk of bias than RCT or prospectively collected data, as equivalent quality controls may not have been implemented. However, archival clinical practice data such as that of NHS Talking Therapies services is an available source of outcome data on a scale seldom seen in psychological therapy research, with treatments implemented with a high degree of standard training and supervision, and this may allow researchers to conveniently address the issue of sample size. More recently, mental health researchers have advocated the use of large electronic health records to optimise clinical prediction models, in view of the sample size limitations of typical clinical trials and the challenges related to data harmonization across clinical trial datasets, which often leads to sparse predictors (Delgadillo & Lutz, 2020; Kessler & Luedtke, 2021). Further, if the aim is to develop a prediction model for use in a particular mental health service, then using data from that same context may boost ecological validity and generalisability. Vieira et al. (2022) comment that using larger, more heterogeneous, naturalistic datasets may produce models with lower prediction accuracy but greater generalisability. Conversely, the finding that trauma related variables may be better predictors of outcome than demographic data presents a problem as many mental health services do not routinely collect this sort of data.

It is important to highlight that PROBAST was not developed to assess ML studies specifically. Some argue that PROBAST may assess ML studies too harshly (Meehan et al., 2022), and others caution that ML methods may be at greater risk of bias under some conditions (Moons et al., 2019; van der Ploeg et al., 2014). Some important features of ML are not assessed by PROBAST, such as hyperparameter selection, which was not reported by seven out of the seventeen studies in this review and can lead to overfitting if performed inappropriately (Delgadillo & Atzil-Slonim, 2022). The inconsistent reporting and application of ML methods identified by this review reiterates the call for specific guidelines and risk of bias assessment tools (Meehan et al., 2022; Vieira et al., 2022), which were under development at the time of the review (Collins et al., 2021).

2.4.2 Sample Size

The finding that none of the studies reported a sample size calculation is congruent with similar reviews of clinical prediction modelling with ML methods (Aafjes-van Doorn et al., 2021; Balki et al., 2019). Determining an appropriate sample size for the development of a clinical prediction model using ML methods is a complex task that depends on several factors. Riley et al. (2020) recently published guidelines for estimating the required sample size that go beyond outcome events per variable (EPV) and other rules of thumb. However, the appropriate sample size also varies according to the particular machine learning method, with some methods requiring larger samples to develop stable models (Dalmaijer et al., 2022; Giesemann et al., 2023; Riley et al., 2021; van der Ploeg et al., 2014). A commonly applied rule-of-thumb is that a minimum of ten EPV is required. However, this is contentious as it is not based on empirical reasoning, and Moons et al. (2019) suggest that an EPV of 20 may be more robust. More precisely it is the number of variable parameters in the model that is of interest (i.e., dummy coded categories and interactions between variables each require the estimation of additional parameters), and when the outcome is categorical the number of outcome events refers to the number of participants in the smallest outcome category. Many studies in this review did not explicitly report the number of candidate predictor variables tested, and where they did it was unclear whether they were reporting the number of variables or number of parameters.

Notably, the four studies that used neuroimaging data did not report the number of candidate predictor parameters. Analysing neuroimaging data typically requires estimation of a large number of parameters, and therefore a large number of participants with the outcome. However, Zhutovsky et al. (2019) and Etkin et al. (2019) had the two smallest samples in the review, and Etkin et al. (2019) did not report the number of participants with the outcome. All four neuroimaging studies identified regions of the brain significantly associated with PTSD treatment response, but there was little consensus among them, and none were externally validated. Etkin et al. (2019) and Zhutovsky et al. (2019) reported accuracies > 80%, but given the issues outlined above this was likely due to overfitting. Similarly, Vieira et al. (2022) found that studies that used neuroimaging data to predict CBT outcomes reported higher accuracy but had smaller sample sizes, suggesting that the higher estimates of accuracy were due to overfitting. Collecting neuroimaging data is typically more expensive and time consuming than collecting questionnaire or patient health record data, which makes the acquisition of an appropriate sample size to analyse high dimensional neuroimaging data even more challenging. Further, this raises doubts about the feasibility of implementing clinical prediction models that require this type of data at scale, particularly in publicly funded health services such as the NHS.

2.4.3 External Validation

The finding that only one study (Herzog et al., 2021) employed external validation procedures mirrors recent reviews of prediction modelling in clinical psychology (Aafjes-van Doorn, 2021; Chekroud et al., 2021; Meehan et al., 2022; Vieira et al., 2022). Moreover, this study only externally validated the model in a randomly partitioned hold-out sample. Some argue that this is not external validation as the training and validation set are subsamples of the same dataset and are likely to be highly correlated and provide overestimates of model performance (Aafjes-van Doorn et al., 2021; Steyerberg, 2019). If possible, splitting the data by time (temporal validation) or geographic location (geographic validation) is a more stringent test of external validity than splitting the data at random (Steyerberg, 2019). Further, some studies had the opportunity to externally validate a model in an independent sample, but replicated model fitting and reported the statistical significance of predictors instead of evaluating model performance metrics (López-Castro et al., 2021; Zhang et al., 2021). This suggests a reluctance among some psychological therapy researchers to shift from seeking to explain relationships between variables to developing pragmatic prediction models (Yarkoni & Westfall, 2017).

2.4.4 Evaluating Model Performance

Nine studies did not report model performance evaluation metrics, including two that applied internal cross-validation (Keefe et al., 2018; Nixon et al., 2021), and none of the studies examined calibration. Therefore, it is unclear how efficacious these models are at predicting therapy outcomes for patients with PTSD. Further, only two studies compared the performance of ML methods to traditional statistical methods: Held et al. (2022) found that five different ML models outperformed logistic regression, but Stuke et al. (2021) found that ordinary linear regression performed slightly better than ADABoost (an ensemble decision tree method). Therefore, it is unclear whether ML methods offer an advantage over traditional statistical methods. Some ML methods may be better than others at predicting treatment outcomes, but only two studies compared the performance of more than one ML method (Etkin et al., 2019; Held et al., 2022).

Four studies applied supervised ML methods to develop a prediction model, but then entered the predictors into a simpler statistical model to estimate parameters and predict outcomes, thereby forgoing any potential advantages of the ML model. López-Castro et al., (2021) commented that variables selected by random forest were not all significant predictors in Poisson regression and suggest that this may be due to correlation with other variables (multicollinearity). However, Poisson regression also makes assumptions about the distribution of the data and the shape of the relationship between the predictor and outcome variables that random forest does not (Mushagalusa et al., 2022).

2.4.5 Recommendations for Future Studies

To properly investigate the potential for ML methods to improve individual prediction of psychological therapy outcomes for PTSD, it is recommended that future studies demonstrate adherence to the ML pipeline in the following ways: [1] Perform a sample size calculation and acquire a large enough dataset; [2] Perform multiple imputation of missing data (stratified by treatment group; Y. Zhang et al., 2021) and report data pre-processing in detail; [3] report all hyperparameter setting (using automated grid search or values selected *a priori*); [4] Apply internal cross-validation during model development and testing; [5] Externally validate (don't repeat model fitting) in an independent sample (temporal or geographic validation are a more stringent test than a random partition; Steyerberg, 2019), and evaluate accuracy, error, discrimination, and calibration. Additionally, it is recommended that studies compare the performance of multiple ML methods against one another and against that of the simplest comparable method (e.g., linear regression or logistic regression). Therefore, the *planning* of future studies should entail completion of the full pipeline to increase the trustworthiness and generalisability of findings.

If ML methods are applied in samples that are too small, with no internal cross-validation, and manual hyperparameter tuning, then it is likely that the model will be overfit to the training data and estimates of model performance will be over-optimistic. Without external validation and calibration, the extent of the optimism and whether the model will generalise is unknown. A recent meta-analysis of ML models found a negative association between study quality and estimates of prediction accuracy, suggesting that poorer quality studies overestimate accuracy (Sajjadian et al., 2021). It is worth noting that for a prediction model to be clinically useful, the model's prediction accuracy does not necessarily need to be high, only better than expert clinical judgement (Ægisdóttir et al., 2006; Cearns et al., 2019). This can be tested in a prospective randomised trial once the external validity of a prediction model has been established (e.g., Delgadillo et al., 2022).

2.4.6 Limitations

ML is an umbrella term that encompasses a broad range of methods, and studies do not always use the term "machine learning". Efforts were made to perform as wide a search as possible, nonetheless it is possible that some relevant studies were not found. Further, the distinction between ML and other statistical methods is not clearly defined, and it is possible that some methods included in this review would not be considered ML by some, and vice versa (Bi et al., 2019). In line with the pre-registration, only studies published in peer reviewed journals were included. This is common practice in psychological therapy reviews, aids replicability of the search procedures, and reduces the likelihood of inclusion of poor-quality studies (Aafjes-van Doorn et al., 2021). However, some relevant studies may have been excluded for this reason (e.g., Cohen, 2018). This review focussed specifically on the prediction of outcome from pre-treatment or baseline data, in the interest of applying ML methods to predict the optimal treatment for individual patients. However, there are other ways that the application of ML methods could potentially improve PTSD treatment outcomes, for example by providing personalised outcome feedback and recommendations during treatment (Bone et al., 2021; Lutz et al., 2019). EndNote 20 reference management software was used to organise and screen search results, and *citationchaser* (Haddaway, 2021) used to conduct forward and backward citation searches. However, use of AI assisted systematic review tools, such as Rayyan (https://www.rayyan.ai/) and Covidence (https://www.covidence.org/), may have increased efficiency and accuracy of searching and screening.

2.4.7 Clinical Implications

This systematic review highlights the need for clinicians to critically evaluate clinical prediction models developed using ML before applying the recommendations of such a model in practice. In particular, clinicians should consider the sample size, the level of evidence (indicated by the presence of internal and/or external cross-validation procedures), and assessments of calibration and discrimination (Delgadillo & Atzil-Slonim, 2022; Steyerberg, 2019).

4.8. Conclusion

Due to the methodological limitations and omissions of the studies identified by this systematic review, it is unclear whether ML methods offer any advantages over traditional statistical methods at predicting psychological therapy outcomes for PTSD. In particular, studies neglected to recruit a sample of an appropriate size informed by a sample size calculation, report hyperparameter setting, perform internal and external cross-validation, and assess model calibration. ML methods have the potential to improve the prediction of treatment outcomes for PTSD, but in order to test this potential they need to be applied rigorously and compared to traditional statistical methods.

CHAPTER 3

External Validation of a Personalised Advantage Index for Patients with Post-Traumatic Stress Disorder

3.1 Introduction

The systematic review reported in Chapter 2 identified a lack of external validation among studies that used machine learning (ML) methods to predict the outcome of psychological therapies for post-traumatic stress disorder (PTSD). Therefore, the aim of this study was to externally validate the personalised advantage index (PAI) model developed by Deisenhofer et al. (2018) in a statistically independent sample of patients treated for PTSD in routine practice at NHS Talking Therapies services. As the dataset used to develop this model did not contain a measure of PTSD, a depression measure was employed as a proxy indicator outcome measure. The secondary aim of this study was to test whether the model developed by Deisenhofer et al. (2018) generalised to a measure of PTSD symptoms. The pre-registered research questions were: [1] Does the model generalise to new independent outcome data? [2] Does the model generalise to a measure of PTSD symptoms?

3.2 Method

3.2.1 Pre-registration

The background, aims, and methodology for this study were pre-registered with As Predicted and the pre-registration can be accessed here: https://aspredicted.org/ca9u5.pdf

3.2.2 Participants, Setting and Interventions

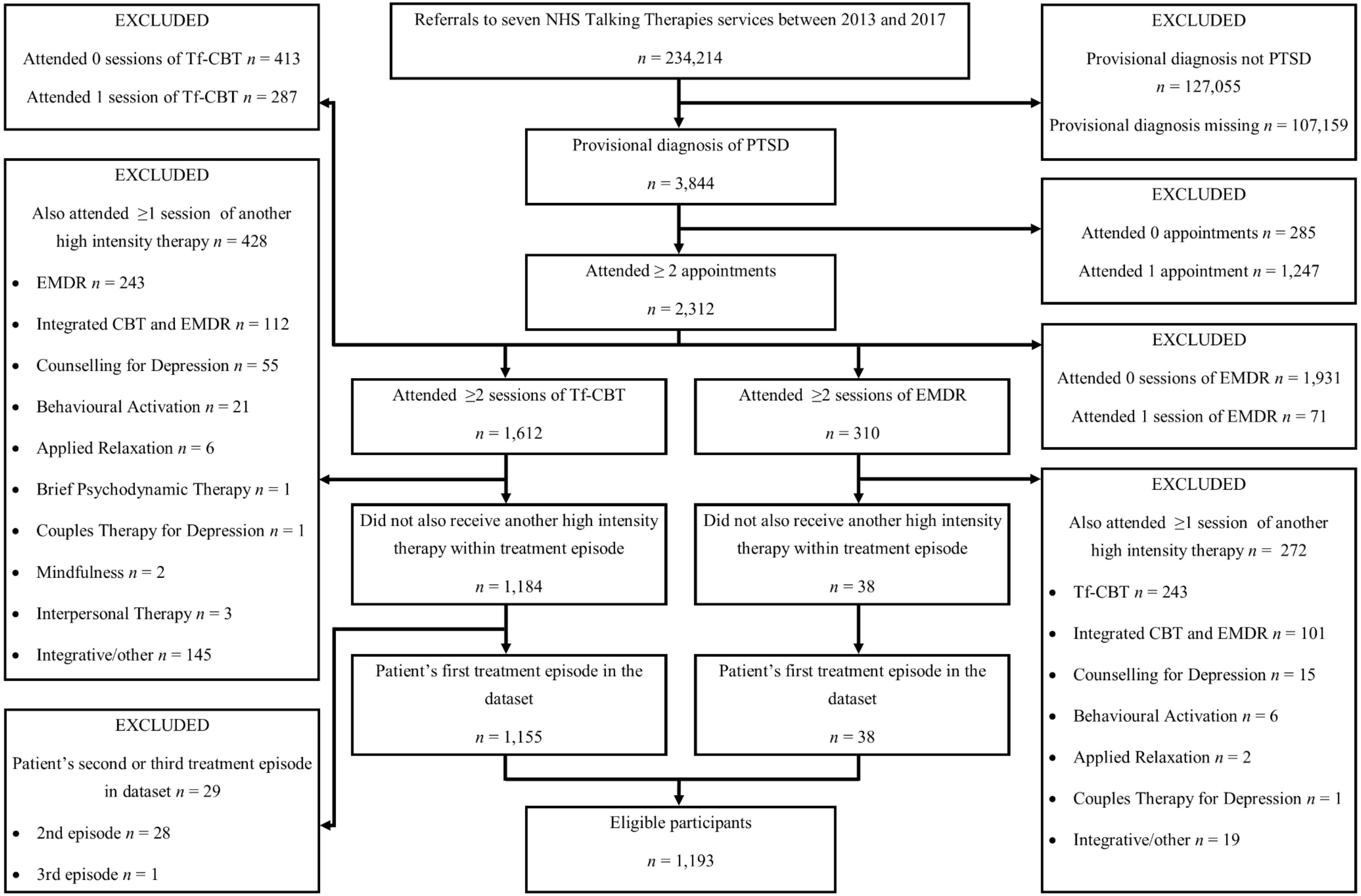
The data used in this study included anonymised clinical records of patients with PTSD accessing NHS Talking Therapies services across seven sites in England between January 2013 and December 2018. Consistent with national clinical guidelines (National

Institute for Health and Care Excellence [NICE], 2018b), patients were allocated to one of two evidence-based psychotherapies for PTSD: Trauma-focussed cognitive behavioural therapy (Tf-CBT; Ehlers et al., 2005) or eye-movement desensitisation and reprocessing (EMDR; Shapiro, 2001). Treatment allocation was based on patient preference and shared decision-making with assessing clinicians. Treatments were delivered by high intensity therapists with the relevant, professionally accredited post-graduate training, practicing independently in outpatient settings under regular clinical supervision. Some patients also accessed brief low intensity interventions (e.g., CBT guided self-help) prior to commencing Tf-CBT or EMDR.

To be included in the study sample, patients were required to have a provisional ICD-10 diagnosis of PTSD (WHO, 2019), and have received ≥ 2 sessions of either Tf-CBT or EMDR (to provide pre- and post-treatment outcome measures). Patients who received more than one high intensity psychological therapy within a treatment episode were excluded to allow for evaluation of the models' treatment-specific outcome predictions. Where the same patient had multiple eligible treatment episodes within the dataset, the first episode was included in the sample and subsequent treatment episodes were excluded, given our interest in the adequacy of timely and accurate treatment selection. Total eligible study sample *N* = 1,193, comprising *n* = 1,155 patients who received Tf-CBT and *n* = 38 patients who received EMDR. The sample selection process is detailed in the STROBE diagram in Figure 3.1.

Figure 3.1

STROBE Flow Diagram Depicting the Sample Selection Process


Referrals to seven NHS Talking Therapies services between 2013 and 2017
n = 234,214
EXCLUDED
Provisional diagnosis not PTSD *n* = 127,055
Provisional diagnosis missing *n* = 107,159
Provisional diagnosis of PTSD
n = 3,844
EXCLUDED
Attended 0 appointments *n* = 285
Attended 1 appointment *n* = 1,247
Attended ≥ 2 appointments
n = 2,312
EXCLUDED
Attended 0 sessions of Tf-CBT *n* = 413
Attended 1 session of Tf-CBT *n* = 287
EXCLUDED
Attended 0 sessions of EMDR *n* = 1,931
Attended 1 session of EMDR *n* = 71
Attended ≥2 sessions of Tf-CBT
n = 1,612
Attended ≥2 sessions of EMDR
n = 310
EXCLUDED
Also attended ≥1 session of another high intensity therapy *n* = 428
• EMDR *n* = 243
• Integrated CBT and EMDR *n* = 112
• Counselling for Depression *n* = 55
• Behavioural Activation *n* = 21
• Applied Relaxation *n* = 6
• Brief Psychodynamic Therapy *n* = 1
• Couples Therapy for Depression *n* = 1
• Mindfulness *n* = 2
• Interpersonal Therapy *n* = 3
• Integrative/other *n* = 145
EXCLUDED
Also attended ≥1 session of another high intensity therapy *n* = 272
• Tf-CBT *n* = 243
• Integrated CBT and EMDR *n* = 101
• Counselling for Depression *n* = 15
• Behavioural Activation *n* = 6
• Applied Relaxation *n* = 2
• Couples Therapy for Depression *n* = 1
• Integrative/other *n* = 19
Did not also receive another high intensity therapy within treatment episode
n = 1,184
Did not also receive another high intensity therapy within treatment episode
n = 38
EXCLUDED
Patient's second or third treatment episode in dataset *n* = 29
• 2nd episode *n* = 28
• 3rd episode *n* = 1
Patient's first treatment episode in the dataset
n = 1,155
Patient's first treatment episode in the dataset
n = 38
Eligible participants
n = 1,193

Note. CBT = cognitive behavioural therapy; EMDR = eye movement desensitisation and reprocessing; NHS = National Health Service; PTSD = post-traumatic stress disorder; Tf-CBT = trauma-focussed cognitive behavioural therapy.

3.2.3 Ethical Approval

Ethical approval was granted by the North West - Greater Manchester West Research Ethics Committee (Ref: 18/NW/0372) for this data source to be used for research. All patients in this dataset provided documented verbal consent for their anonymised data to be used for research.

3.2.4 Measures

3.2.4.1 Psychometric Measures

Due to the absence of a measure of PTSD symptoms in the model development dataset Deisenhofer et al. (2018) used the Patient Health Questionnaire 9 (PHQ-9; Kroenke et al., 2001) as the primary outcome measure, and this was replicated in the current study. The PHQ-9 is a validated, nine-item, self-report measure of depression severity, based on the DSM-IV criteria for major depression. Each of the 9 items describes a symptom of depression, and patients rate each item on a scale from 0 (*not at all*) to 3 (*nearly every day*). Total PHQ-9 scores range from 0 to 27, with higher scores indicating greater number and frequency of depression symptoms. Kroenke et al. (2001) recommended classifying scores ≥ 10 as moderate depression, and reported sensitivity of 88% and specificity of 88%. Kroenke et al. (2001) reported good reliability (Cronbach's a = 0.89) in a primary care sample. A change of ≥ 6 points on the PHQ-9 has been recommended as an index of reliable improvement or deterioration in symptoms (Richards & Borglin, 2011).

The current dataset contained a self-report measure of PTSD symptoms, the Impact of Event Scale-Revised (IES-R; Weiss, 2007), but the high proportion of missing values (85.8% pre-treatment, 87% post-treatment) precluded investigation of the second pre-registered research question. However, there was a significant, medium-sized, positive correlation between pre-treatment PHQ-9 and IES-R score (r (176) = .44, p < .001), and a significant, large, positive correlation between post-treatment PHQ-9 and IES-R score (*r* (160) = .75, *p* < .001).

The PHQ-9 was administered before every session, along with the Generalised Anxiety Disorder 7 (GAD-7; Spitzer et al., 2006) and the Work and Social Adjustment Scale (WSAS; Mundt et al., 2002). The GAD-7 is a validated seven-item self-report measure of anxiety symptoms, scores range from 0-21 with higher scores indicating more severe symptoms. The WSAS is a validated five-item self-report measure of the extent to which a person's mental health problems impair their daily functioning. WSAS scores range from 0-40 with higher scores indicating greater functional impairment. Pre-treatment scores were extracted from each patient's first high intensity treatment session, and post-treatment scores extracted from their last high intensity treatment session.

3.2.4.2 Demographic and Health Variables

Age, gender, ethnicity, disability, and long-term condition (LTC) data were extracted from patient records. For some patients with multiple referrals in the dataset age data was only available for the most recent referral; age was calculated for earlier referrals by subtracting the number of years between referral dates from the patient's age at the most recent referral. Ethnicity was based on the Office for National Statistics ethnic categories (Office for National Statistics, n.d.) and was self-reported. Disability was a binary indicator of the presence of any patient-reported disability. LTC was a binary indicator of whether a patient had a long-term physical health condition such as diabetes, arthritis, or a chronic respiratory condition.

Employment status consisted of eight categories: 'employed', 'unemployed', 'student', 'long-term sick or disabled', 'homemaker', 'not receiving benefits and not seeking work', 'unpaid voluntary work and not seeking work', and 'retired'. Medication status referred to antidepressant medication and consisted of three categories: 'prescribed and taking', 'prescribed but not taking', and 'not prescribed'. Employment and medication status were self-reported and were recorded at every appointment; pre-treatment values were extracted from each case's first high intensity treatment session.

3.2.5 Pre-processing of Data

3.2.5.1 Missing Data and Multiple Imputation

See Appendix G for the proportion of missing values on each variable. In the EMDR group, age, gender, disability, pre-treatment PHQ-9 and pre-treatment GAD-7 had no missing values; but in the Tf-CBT group only age and disability had no missing values. Six variables had > 5% missing values in both treatment groups. These were LTC (Tf-CBT = 38.8%, EMDR = 34.2%), post-treatment PHQ-9 (Tf-CBT = 6.1%, EMDR = 7.9%), post-treatment GAD-7 (Tf-CBT = 6.1%, EMDR = 7.9%), pre-treatment WSAS (Tf-CBT = 13.3%, EMDR = 23.7%), post-treatment WSAS (Tf-CBT = 20.1%, EMDR = 23.7%), and medication (Tf-CBT = 19.4%, EMDR = 10.5%). Little's missing completely at random (MCAR) test was non-significant (X^2 (52) = 68.99, p = .057), suggesting that missing values were MCAR.

Multiple imputation of missing values was performed via a non-parametric random forest method, using the missForest package in R (Stekhoven & Bühlmann, 2012). This method was applied instead of the pre-registered method of multiple imputation by chained equations to replicate the methods used by Deisenhofer et al. (2018). There is evidence that random forest produces smaller imputation error than multiple imputation by chained equations (Waljee et al., 2013). For each variable with missing values, missForest trains a random forest prediction model on the non-missing values with all other variables as predictors, and then uses it to impute the missing values. This is done for both continuous and categorical variables. For a detailed explanation of random forests, see Breiman (2001a).

Deisenhofer et al. (2018) applied missForest to 15 variables, including the continuous variables age, pre-treatment PHQ-9 score, post-treatment PHQ-9 score, pre-treatment GAD-7 score, post-treatment GAD-7 score, pre-treatment WSAS score, and post-treatment WSAS score; and the categorical variables gender (binary), disability (binary), long-term physical health condition (binary), ethnicity (five factors), pre-treatment employment status (eight factors), and pre-treatment medication usage (three factors). Additionally, a binary 'treatment' variable (Tf-CBT or EMDR) was entered as a categorical variable, and a numeric 'group' variable (1 = EMDR, 0 = Tf-CBT) was entered as a continuous variable. Default settings were used. missForest was applied to the same variables used by Deisenhofer et al. (2018) except for the indicators of treatment group; instead multiple imputation was performed for each treatment group independently, in line with recent recommendations (J. Zhang et al., 2023).

The missForest method produces out-of-bag (OOB) error estimates in the form of Normalised Root Mean Squared Error (NRMSE) for continuous variables, and Proportion Falsely Classified (PFC) for categorical variables. Both PFC and NRMSE can be interpreted as the proportion of values incorrectly imputed, therefore a value close to 0 indicates good performance and a value close to 1 indicates poor performance. In the current sample, missForest produced NRMSE = .33 and PFC = .32 for the Tf-CBT group (*N* = 1,155), and NRMSE = .33 and PFC = .23 for the EMDR group (*N* = 38). Imputed values on continuous variables were rounded to the nearest integer. No imputed values were out of range.

3.2.5.2 Propensity Score Matching

As this study was based on routinely collected clinical data, patients were not randomly allocated to treatment. Treatment selection was based on a shared decision-making process between the patient and clinician and may have been influenced by patient characteristics. As such, it is possible that there are systematic differences in the pre-treatment characteristics of the two treatment groups, which could confound the relationship between treatment and outcome. This is known as confounding by indication (Kyriacou & Lewis, 2016).

Propensity score matching (PSM; Rosenbaum & Rubin, 1983) is a commonly used method to control for confounding by indication and is recommended for personalised treatment selection studies using routine clinical data (Kessler et al., 2019). A propensity score is a patient's probability of being allocated to a particular treatment group based on their observed pre-treatment characteristics and is typically estimated via logistic regression. Patients in the comparator group are then selected for inclusion in the sample by the similarity of their propensity score to that of patients in the treatment group (i.e., patients are matched between groups based on their propensity score). Rosenbaum and Rubin (1983) demonstrated that patients with similar propensity scores have comparable distributions of the observed covariates. In this way, PSM produces a balance of observed pre-treatment covariates like that produced by randomisation.

Patients who received Tf-CBT were propensity-score matched to the *n* = 38 patients who received EMDR at a ratio of 3:1, producing a Tf-CBT group of *n* = 114. The ratio of 3:1 reflects the relative infrequency of routine service delivery of EMDR, due to the small workforce of qualified EMDR practitioners. The resulting study sample size of *N* = 152 is smaller than the pre-registered sample size of *N* = 180. Given the size of the initial dataset (*N* = 234,214 referrals) it was expected that there would be more cases who accessed protocol driven EMDR for PTSD, but EMDR was most often delivered as part of an integrated treatment with Tf-CBT (see Figure 3.1).

PSM was performed using the MatchIt package in R (Stuart et al., 2011), applying the *optimal matching* method. The variables used for PSM were the same nine variables used by Deisenhofer et al. (2018): gender, age, LTC, disability, employment, medication, and pre-treatment scores on the PHQ-9, GAD-7, and WSAS. The standardised mean difference (SMD) method was used to assess the difference between groups on each variable, whereby an SMD < .25 is considered an adequate match between groups. Prior to PSM, pre-treatment WSAS score and employment status had an SMD > .25 indicating baseline differences between the two groups on these two variables. Following PSM all variables had an SMD < .25.

3.2.5.3 Sample Characteristics After Multiple Imputation and PSM

Sample characteristics are presented in Table 3.1. After matching, the current external validation sample was significantly different to the model development sample (Deisenhofer et al., 2018) on two variables: pre-treatment WSAS score and disability. When tested with Welch's t-test, pre-treatment WSAS in the validation sample (N = 152, mean = 17.07, SD = 9.72) was significantly lower than in the development sample (N = 225, mean = 21.13, SD = 10.28; t (335.81) = 3.89, p < .001); pre-treatment WSAS in the Tf-CBT group of the validation sample (n = 114, mean = 17.13, SD = 10.01) was significantly lower than the Tf-CBT group in the development sample (n = 150, mean = 21.10, SD = 10.02; t (243.48) = 3.20, p < .002); and pre-treatment WSAS in the EMDR group of the validation sample (n = 38, mean = 16.87, SD = 8.95) was significantly lower than the EMDR group in the development sample (n = 75, mean = 21.18, SD = 10.84; t (88.13) = 2.25, p < .027). Chi-square tests indicated a significantly lower rate of disability in the validation sample (19.74%) than the development sample (48%; X^2 (1) = 30.02, p < .001); a significantly lower rate of disability in the Tf-CBT group of the validation sample (21.05%) than the Tf-CBT group of the development sample (50.67%; X^2 (1) = 22.90, p < .001); and a significantly lower rate of disability in the EMDR group of the validation sample (15.79%) than the

EMDR group of the development sample (42.67%; X^2 (1) = 7.00, p = .008).

Table 3.1

Sample Characteristics (After Multiple Imputation)

	Tf-CBT	Tf-CBT after PSM	EMDR
	(*N* = 1155)	(*N* = 114)	(*N* = 38)
	Mean (SD) or %	Mean (SD) or %	Mean (SD) or %
PHQ-9 pre	16.05 (6.18)	14.93 (6.62)	14.82 (6.43)
PHQ-9 post	11.43 (7.58)	10.37 (7.82)	9.05 (7.17)
GAD-7 pre	14.93 (4.81)	14.37 (5.27)	14.16 (5.00)
GAD-7 post	10.62 (6.52)	9.99 (7.09)	8.68 (6.40)
WSAS pre	21.11 (9.67)	17.13 (10.01)	16.87 (8.95)
WSAS post	15.74 (11.01)	13.63 (11.05)	11.16 (8.89)
Gender (female)	62.17%	59.65%	63.16%
Age	38.94 (12.79)	43.02 (14.87)	40.63 (12.50)
LTC	29.09%	30.70%	26.32%
Disability	8.83%	21.05%	15.79%
Employment pre			
Employed	49.44%	47.37%	47.37%
Student	4.24%	0.00%	0.00%
Unemployed	3.55%	3.51%	5.26%
Long-term sick	17.40%	9.65%	7.89%
Other [a]	25.37%	39.47%	39.47%
Medication[b] pre			
Prescribed	55.06%	50.00%	50.00%
Prescribed not taking	3.55%	4.39%	5.26%
Not prescribed	41.39%	45.61%	44.74%
Ethnicity[c]			
White	74.46%	77.19%	89.47%
Mixed/Multiple	3.64%	2.63%	5.26%
Asian/Asian British	7.62%	7.02%	2.63%
Black/Black British	10.22%	9.65%	2.63%
Other	4.07%	3.51%	0.00%

IAPT appointments attended	8.61 (5.15)	9.07 (5.26)	7.00 (4.51)
Hight intensity treatment sessions	7.50 (4.93)	7.88 (5.16)	5.68 (4.53)
Accessed low intensity interventions	65.63%	65.79%	44.74%

Note. EMDR = Eye-movement desensitization and reprocessing; GAD-7 = Generalised Anxiety Disorder 7; LTC = Long-term medical condition; PHQ-9 = Patient Health Questionnaire 9; PSM = Propensity score matching; Tf-CBT = Trauma-focussed cognitive behavioural therapy; WSAS = Work and Social Adjustment Scale.

[a] Employment *Other* = Voluntary work, homemaker, carer, or retired.

[b] Medication = Antidepressant medication.

[c] Ethnicity = Office for National Statistics ethnic group.

3.2.6 Data Analysis Strategy

3.2.6.1 Comparing Tf-CBT and EMDR Treatment Outcomes

Treatment outcomes were compared between the Tf-CBT and EMDR groups by comparing the 95% confidence intervals of the pre-treatment to post-treatment effect size (*d*) on the PHQ-9, GAD-7, and WSAS. If the confidence intervals overlap, this indicates that there was no statistically significant difference between the two groups in pre- to post-treatment change. Effect sizes and confidence intervals were calculated using the method described by Minami et al. (2008), adjusted for non-normal distributions using Spearman's rank correlation (see Appendix H for Q-Q plots and correlation matrices).

3.2.6.2 Predicting Treatment Outcomes

The two linear regression models developed by Deisenhofer et al. (2018) were applied to predict outcomes across both treatment groups (*N* = 152) from patients' pre-treatment scores on the predictor variables selected by the genetic algorithm during model development. Using the stats package in R, each regression model was fitted to the respective training data via the `lm()` function, giving the same coefficients reported by Deisenhofer et al. (2018), and was then used to predict post-treatment PHQ-9 score in the external validation sample via the `predict()` function.

The linear regression equation for the EMDR model was:

$$Y = 8.78 + (0.44 \times \text{PHQ-9 score}) + (4.40 \times \text{Medication status})$$

And the linear regression equation for the Tf-CBT model was:

$$Y = 9.83 + (0.24 \times \text{WSAS score}) - (4.99 \times \text{Employment status}) - (0.10 \times \text{Age}) - (2.09 \times \text{Gender})$$

Deisenhofer et al. (2018) centred continuous baseline variables around the group mean. Baseline PHQ-9 score was centred around the EMDR group mean (15.22), and baseline WSAS was centred around the Tf-CBT group mean (21.11). As such, pre-treatment PHQ-9 and WSAS were centred around these respective values in the whole validation sample. Employment status was reduced to a binary variable, with 'employed' and 'student' coded as 0.5, and 'unemployed', 'long-term sick', and all other categories coded as -0.5. Medication status was reduced to a binary variable, with 'prescribed and taking' and 'prescribed but not taking' coded as 0.5, and 'not prescribed' coded as -0.5.

3.2.6.3 Evaluating Model Performance

To evaluate prediction accuracy, R^2 was calculated by squaring the correlation (Pearson's *r*) between the observed post-treatment PHQ-9 scores and the scores predicted by each model. R^2 can be interpreted as the proportion of variance in treatment outcome explained by the model, with a maximum value of 1 indicating perfect prediction accuracy, and values close to 0 indicating poor prediction accuracy. R^2 was examined for each prediction model in each of the treatment groups, if the two models make treatment-specific predictions, then it would be expected that the Tf-CBT model makes more accurate predictions in the Tf-CBT group than in the EMDR group, and the EMDR model makes more accurate predictions in the EMDR group than in the Tf-CBT group. To evaluate model prediction error, Root Mean Squared Error (RMSE) was calculated by taking the square root of the mean squared differences between the predicted and observed scores. Lower RMSE values indicate less prediction error, and higher values indicate more prediction error.

R^2 and RMSE estimates in the external validation sample were compared to apparent R^2 and RMSE estimates in the model development sample (i.e., without the LOO cross-validation that was applied to internally validate the model during model development). Additionally, for comparison with Deisenhofer et al. (2018), *true error* was calculated as the mean absolute difference between the observed post-treatment PHQ-9 score and the factual predictions (i.e., Tf-CBT model predictions in the Tf-CBT group, and EMDR model predictions in the EMDR group).

3.2.6.4 Comparing Model-Indicated Optimal and Suboptimal Treatment Outcomes

The model-indicated optimal treatment was identified for each patient by comparing their Tf-CBT and EMDR model predictions; the treatment with the lowest predicted post-treatment PHQ-9 score was labelled their optimal treatment, and the treatment with the highest predicted post-treatment PHQ-9 score was labelled their suboptimal treatment. Patients were then grouped by whether they had received their optimal or suboptimal treatment, and average treatment outcomes were compared between the two groups. Patients were labelled as having reported reliable change in symptoms if their post-treatment PHQ-9 score was ≥ 6 points lower than their pre-treatment PHQ-9 score (Richards & Borglin, 2011). The rate of reliable improvement was compared between the optimal and suboptimal treatment groups with a chi-square test. For comparison with Deisenhofer et al. (2018), Number Needed to Treat (NNT) was estimated using the effect size calculator provided by

Lenhard and Lenhard (2016).

3.2.6.5 Personalised Advantage Index

The PAI was calculated by subtracting the predicted outcome of each patient's optimal treatment from the predicted outcome of their suboptimal treatment. In this way, the PAI represents the predicted difference in outcome between the optimal and suboptimal treatment for each patient; the greater the PAI value, the more likely that patient is to benefit from receiving their optimal treatment rather than their suboptimal treatment. Patients with a PAI ≥ 1 standard deviation are those most likely to benefit from personalised treatment selection. The standard deviation of PAI scores in the development sample was 1.92. As a test of the clinical utility of the PAI, a regression analysis was performed predicting post-treatment PHQ-9 score from a binary indicator of whether a patient received their optimal or suboptimal treatment, among patients with a PAI ≥ 1 standard deviation (i.e., excluding patients with a PAI < 1.92 from the analysis). Pre-treatment PHQ-9 score was included as a covariate to control for baseline symptom severity, and propensity score was included as a covariate as a secondary control (after PSM) for confounding by indication (D’Agostino, 1998).

3.3 Results

Results follow the structure set out in the data analysis strategy (section 3.2.6).

3.3.1 Comparison of average treatment effect

The median number of EMDR sessions was 4 (inter-quartile range = 2 – 8.25) and the mode was 2 (range = 2 – 20). The median number of Tf-CBT sessions was 6 (inter-quartile range = 4 – 10) and the mode was 6 (range = 2 – 29). Pre- to post-treatment effect sizes (*d*) for the whole matched sample, the matched Tf-CBT group, and the EMDR group, are presented in Table 3.2. Comparison of the 95% confidence intervals suggest no significant difference in treatment effect size between groups. Of the current sample (which excluded cases who only received one session of therapy), 40.13% (*n* = 61 / 152) reported a reliable improvement in depression symptoms. This included 50% of the EMDR group (*n* = 19 / 38) and 36.84% of the Tf-CBT group (*n* = 42 / 114). The difference in these rates of reliable improvement was not statistically significant (X^2 (1) = 1.54, *p* = .214).

Table 3.2

Treatment Outcome in the Total Sample and Matched Tf-CBT and EMDR Groups

Measure	Sample	*d*	95% Confidence Interval	
			Lower limit	Upper limit
PHQ-9	Total (N = 152)	0.74	0.58	0.90
	Tf-CBT (n = 114)	0.88	0.56	1.20
	EMDR (n = 38)	0.68	0.50	0.87
GAD-7	Total (N = 152)	0.89	0.71	1.07
	Tf-CBT (n = 114)	1.07	0.69	1.45
	EMDR (n = 38)	0.83	0.62	1.03
WSAS	Total (N = 152)	0.41	0.27	0.56
	Tf-CBT (n = 114)	0.62	0.32	0.93
	EMDR (n = 38)	0.35	0.18	0.52

Note. EMDR = Eye-movement desensitization and reprocessing; GAD-7 = Generalised Anxiety Disorder 7; PHQ-9 = Patient Health Questionnaire 9; Tf-CBT = Trauma-focussed cognitive behavioural therapy; WSAS = Work and Social Adjustment Scale.

Effect sizes and confidence intervals calculated using the method described by Minami et al. (2008).

3.3.2 Model Evaluation

R^2 and RMSE for the Tf-CBT model and the EMDR model in the whole sample, Tf-CBT group, and EMDR group, of the development sample and external validation sample are presented in Table 3.3. The R^2 values indicate that although each prediction model demonstrates better predictive accuracy in the corresponding treatment group in the development sample, this was not the case in the validation sample. The pattern of results in the development sample is what would be expected if the two models make treatment specific predictions: The Tf-CBT model predicts Tf-CBT outcomes with a greater accuracy than EMDR outcomes, and with greater accuracy than the EMDR model; and the EMDR model predicts EMDR outcomes with greater accuracy than Tf-CBT outcomes, and with greater accuracy than the Tf-CBT model. However, this pattern did not replicate in the external validation sample, suggesting that these two models make general prognostic predictions that are not treatment specific.

Table 3.3

*Model Prediction Accuracy (*R^2*) and Error (RMSE) of the Tf-CBT and EMDR Prediction Models in the Development and Validation Samples*

Sample	Tf-CBT Model		EMDR Model	
	R^2	RMSE	*R*^{2}	RMSE
Development sample (*N* = 225)	.28	6.45	.20	6.84
Development Tf-CBT (*n* = 150)	.38	5.97	.14	7.21
Development EMDR (*n* = 75)	.11	7.32	.35	6.05
Validation sample (*N* = 152)	.30	6.62	.45	5.85
Validation Tf-CBT (*n* = 114)	.28	6.92	.47	6.00
Validation EMDR (*n* = 38)	.38	5.64	.42	5.40

Note. EMDR = Eye-movement desensitization and reprocessing; PHQ-9 = Patient Health Questionnaire 9; RMSE = Root Mean Squared Error; Tf-CBT = Trauma-focussed cognitive behavioural therapy.

R^2 was calculated by squaring the correlation (Pearson's *r*) between the predicted and observed post-treatment PHQ-9 scores and can be interpreted as the proportion of variance explained by the model.

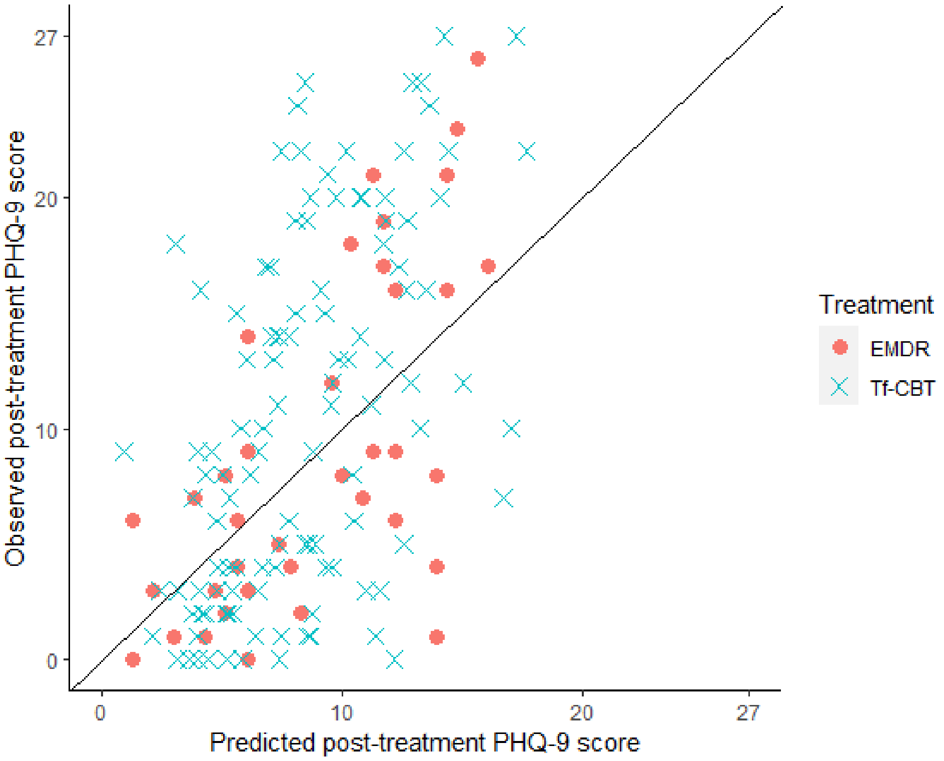
RMSE was calculated by taking the square root of the mean of the squared differences between the predicted and observed post-treatment PHQ-9 scores.

True error for the whole sample was 5.44, compared to 5.07 in the development sample with LOO cross-validation, and 4.83 without. For the Tf-CBT group, true error was 5.76, compared to 5.37 in the development sample with LOO cross-validation, and 4.74 without. For the EMDR group, true error was 4.49, compared to 4.92 in the development sample with LOO cross-validation, and 5.03 without.

The calibration plot presented in Figure 3.2 plots the observed final session PHQ-9 scores against the factual predictions made by each of the regression models. The closer the points are to the diagonal line, the more accurate the prediction. The calibration plot suggests that the models make more accurate predictions for low scores, with more error in predictions for higher scores, and neither model predicts any scores at the higher end of the scale.

Figure 3.2

Calibration Plot Comparing Predicted and Observed Post-treatment PHQ-9 Scores



Note. EMDR = Eye-movement desensitization and reprocessing; PHQ-9 = Patient Health Questionnaire 9; Tf-CBT = Trauma-focussed cognitive behavioural therapy.

3.3.3 Predicting Optimal Treatment

Of the whole sample, 57.23% (*n* = 87) received their model-indicated optimal treatment, including 47.37% (*n* = 18) of the EMDR group, and 60.53% (*n* = 69) of the Tf-CBT group. There was no significant difference between the Tf-CBT and EMDR group in the number of cases who received their model-indicated optimal treatment (X^2 (1) = 1.51, *p* = .219). The mean observed post-treatment PHQ-9 score for the optimal treatment group was 11.09 (SD = 7.37), and for suboptimal treatment group the mean was 8.63 (SD = 7.89). This is a mean difference of 2.46, corresponding to a Cohen's *d* = .32 (95% CI [0.00, 0.65]). However, a similar average group difference in PHQ-9 score was observed pre-treatment: The mean observed pre-treatment PHQ-9 score for the optimal group was 15.92 (SD = 5.72), and for the suboptimal group was 13.54 (SD = 7.34), mean difference = 2.38, Cohen's *d* = .37 (95% CI [0.04, 0.69]). Hence, it is necessary to control for differences in baseline symptom severity when comparing average treatment effect, as follows in the regression analysis. Rates of reliable improvement are presented in Table 3.4. There was no significant difference in the rate of reliable improvement between patients who received their optimal treatment (39.08%) versus those who received their suboptimal treatment (41.54%; X^2 (1) = 0.02, *p* = 0.890).

Table 3.4

*Comparison of the Rate of Reliable Improvement in PHQ-9 Score Between Patients Who Received Their Model-Indicated Optimal Treatment (*N *= 87) and Patients Who Received Their Model-Indicated Suboptimal Treatment (*N *= 65)*

Treatment Received	Reliable Improvement	
	Yes *n* (%)	No *n* (%)
Optimal	34 (39.08%)	53 (60.92%)
Suboptimal	27 (41.54%)	38 (58.46%)

Note. X^2 (1) = 0.02, *p* = .890

In instances where there is a lower rate of the desired outcome in the treatment group than in the comparator group, NNT becomes the Number Needed to Harm (NNH). There was a 2.46% lower rate of reliable improvement in the optimal treatment group, which corresponds to an NNH = 40.68. This suggests that for every 40 to 41 patients who received their model-indicated optimal treatment, an additional case would not experience reliable improvement, compared to patients who received their model-indicated suboptimal treatment.

3.3.4 The Personalised Advantage Index

The mean PAI score was 2.85 (SD = 1.96), the minimum was 0.03 and maximum was 9.39. In the validation sample, 7.89% (*n* = 12 / 152) had a PAI of 0.5 or less, compared to 14.22% of the development sample. In the validation sample, 61.18% (*n* = 93 / 152) had a PAI greater than or equal to 1.92 (the SD reported by Deisenhofer et al., 2018), 52.69% of whom (*n* = 49 / 93) received their model-indicated optimal treatment, and 41.9% of whom (*n* = 39 / 93) had a reliable change in symptoms. Among the *n* = 93 patients with a PAI ≥ 1.92 there was no significant difference in the rate of reliable improvement between patients who received their optimal versus those who received their suboptimal treatment (X^2 (1) = 0.0004, *p* = .984).

A binary indicator of having received model-indicated optimal treatment was entered into a regression model with post-treatment PHQ-9 score as the dependent variable, and pre-treatment PHQ-9 score and propensity score as covariates. To address heteroscedasticity in the residuals, post-treatment PHQ-9 score was square root transformed and propensity score was log transformed (see Appendix I for further details). The results of the regression analysis presented in Table 3.5 revealed that among the *n* = 93 patients with a PAI ≥ 1.92, receiving model-indicated optimal treatment had no significant effect on post-treatment PHQ-9 score (β = 0.12, *p* = .168), adjusted for pre-treatment PHQ-9 score (β = 0.61, *p* < .001) and propensity score (β = -0.03, *p* = .691).

Table 3.5

*The Effect of Model-Indicated Optimal Treatment on Post-Treatment PHQ-9 Score (Square Root Transformed), Adjusted for Pre-Treatment PHQ-9 Score and Propensity Score Estimate (Log Transformed), Among Patients With a PAI ≥ 1.92 (*N *= 93)*

Effect	B	SE	β	t	95% CI		*p*
					Lower	Upper	
Intercept	0.45	0.51		0.87	-0.57	1.46	.386
Optimal treatment	0.33	0.24	0.12	1.39	-0.14	0.81	.168
PHQ-9 pre	0.13	0.02	0.61	7.23	0.09	0.17	< .001
Propensity score (log)	-0.14	0.35	-0.03	-0.40	-0.84	0.56	.691

Note. F (3, 89) = 22.77, p < .001, R^2 = .43

3.4 Discussion

This study was the first external validation of a PAI for the treatment of PTSD. Two linear regression models, developed using a genetic algorithm, were applied to predict outcomes of Tf-CBT and EMDR in a statistically independent sample. In the model development sample (Deisenhofer et al., 2018), the Tf-CBT model predicted outcomes for the Tf-CBT group with greater accuracy than the EMDR model; and the EMDR model predicted outcomes for the EMDR group with greater accuracy than the Tf-CBT model. This suggests that each model makes treatment-specific outcome predictions. However, in the current external validation sample, this pattern of results was not replicated. This suggests that these models are simply prognostic models that predict PTSD treatment outcome independent of treatment type, and any differential outcome predicted in the model development sample is due to overfitting of the model to the treatment group.

Unlike the model development sample, in the external validation sample there was no significant difference in rates of reliable improvement between patients who received their model-indicated optimal treatment versus those who received their suboptimal treatment. When the clinical utility of the PAI was tested among patients with a robust treatment recommendation (i.e., PAI ≥ 1SD), receiving optimal treatment was not significantly associated with treatment outcomes. The NNH suggested that for every 41 patients treated with their model-indicated optimal treatment, one additional patient would not attain reliable improvement. However, as highlighted by Kraemer and Kupfer (2006), NNT/NNH is unstable when the difference in the rate of outcome between the treatment and comparator group is close to 0, and as the difference is not statistically significant, NNT/NNH could fluctuate between large positive and large negative values in different samples.

These findings suggest that the PAI model developed by Deisenhofer et al. (2018) does not predict differential treatment outcomes for PTSD in independent data collected from a similar setting. This could be attributed to methodological issues in both the training and testing phases, which will be discussed further below. Nevertheless, the results are in line with previous findings on the transferability of some prediction models for treatment selection to external validation data from the same (B. Schwartz et al., 2021) and other comparable studies (Van Bronswijk et al., 2021).

3.4.1 Limitations of the Model Development Method

In a simulation study, Luedtke et al. (2019) found that a minimum *n* = 300 patients per treatment group was required to reliably detect predictors of differential treatment response. The sample used by Deisenhofer et al. (2018) to develop the model was considerably smaller than this, with *n* = 150 patients in the CBT group and *n* = 75 in the EMDR group. Concurrently, genetic regression may not be the best method of developing a model for this task. In genetic regression, a genetic algorithm performs predictor selection, and the model parameters are estimated by ordinary linear regression. It is possible that this method, combined with the small sample size led to overfitting of the model. Deisenhofer et al. (2018) applied LOO cross-validation, but this only adjusts optimism when evaluating prediction accuracy, and does not control overfitting that occurs during predictor selection and parameter estimation (Kessler et al., 2017).

Alternatively, *penalised regression* methods such as *elastic net* control for overfitting during model development by shrinking small coefficients towards zero (Zou & Hastie, 2005). Held et al. (2022) applied six different machine learning algorithms to predict treatment response trajectory in a sample of military veterans who accessed CPT for PTSD, and in a randomly partitioned hold-out sample found that elastic net most accurately predicted minimal response (*gradient boosted models* most accurately predicted fast response). Herzog et al. (2021) used elastic net to predict change in symptoms following psychological therapy for PTSD and found that the model generalised to a hold-out test sample (training R^2 = .17, validation R^2 = .16). Delgadillo and Gonzales Salas Duhne (2020) used elastic net to develop a PAI for two psychological therapies for depression and found a significantly higher rate of reliable improvement for patients who received their model-indicated optimal treatment in a hold-out validation sample. In addition to using elastic net, Delgadillo and Gonzales Salas Duhne (2020) used a much larger sample (N = 1,435), and applied bootstrapping when estimating model parameters, which has been shown to improve external validity (Steyerberg et al., 2003). Bootstrapping may be a more robust method of internal validation than LOO and similar internal cross-validation procedures (Steyerberg et al., 2001), particularly as the N-1 'training sets' (or *folds*) in LOO are unlikely to be substantially different from one another (Hastie et al., 2009).

Development of the PTSD PAI model was likely further limited by the available variables; the dataset did not contain any measures of PTSD symptoms or trauma-related variables. Recent studies have found that clinical, PTSD and trauma related variables are better predictors of PTSD treatment outcome than demographic variables (Held et al., 2022; Herzog et al., 2021; Hoeboer, Oprel, et al., 2021; Keefe et al., 2018; Stuke et al., 2021). In the absence of a measure of PTSD symptoms, the PHQ-9 was used as a proxy outcome measure.

Whilst the PHQ-9 correlates with PTSD severity, it is not clear whether the models predict change in PTSD symptoms, or only change in depressive symptoms; the large proportion of missing values on the IES-R in the validation sample precluded investigation of this. Additionally, pre-treatment PHQ-9 score was selected as a predictor in the EMDR model, but not the Tf-CBT model. Hence, the Tf-CBT model predicts post-treatment PHQ-9 without adjusting for pre-treatment PHQ-9 score, and it is questionable whether this is a valid measure of treatment outcome.

3.4.2 Limitations of the Current Study

In a resampling study, Collins et al. (2016) found that a minimum *n* = 100 was required to obtain reliable estimates of prediction model performance in external validation, and in the current study there were only *n* = 38 patients in the EMDR group. This was because most patients who received EMDR also received at least one session of Tf-CBT, precluding their inclusion in the sample. Recent systematic reviews have found that sample size remains a common limitation of clinical psychology prediction modelling research (Meehan et al., 2022; Vieira et al., 2022).

There was also considerable missing data, up to 38.8% on LTC. LTC was not a predictor in either of the models, but pre-treatment WSAS (predictor in the Tf-CBT model) was missing 13% in the unmatched Tf-CBT group and 23% in the EMDR group, and medication (predictor in the EMDR model) was missing 19% in the unmatched Tf-CBT group and 10% in the EMDR group. This could have introduced additional biases and given the small number of predictors in each of the models any bias in these variables is problematic. But missing data is a common issue in clinical research, and the proportion missing in this study was comparable with that of Van Bronswijk et al. (2021).

Although there is evidence that missForest outperforms multiple imputation with chained equations, it is not without its limitations, and there are other random forest-based imputation methods that may have some advantages over missForest (Hong & Lynn, 2020). The effect of different imputation methods on the accuracy and generalisability of prediction models is yet to be empirically tested. The proportion of missing data on the IES-R measure of PTSD symptoms precluded multiple imputation of this variable and the investigation of the second research question. As the PAI did not generalise to PHQ-9 scores in the external validation sample it appears unlikely that it would have generalised to IES-R scores.

Patients in the model development sample had a significantly higher rate of disability and functional impairment than the validation sample. Differences in samples, including heterogeneity in clinical presentations, treatment delivery, and available predictors can result in poorer model performance in external validation (Hehlmann et al., 2023; Van Bronswijk et al., 2021). This may have reduced the likelihood that the models would generalise to the external validation sample, particularly as pre-treatment WSAS score was a predictor in the Tf-CBT model. However, it could be argued that personalised treatment prediction models need to be robust to varying distributions of covariates if they are to be implemented in clinical practice. There is evidence that prediction models developed using machine learning methods with a sufficient sample size can generalise to samples recruited from different geographic locations (Bone et al., 2021) and at different times (Delgadillo et al., 2020).

Most patients did not receive the NICE (2018b) recommended 8-12 sessions of Tf-CBT or EMDR. Also, as this study used naturalistic data from routine clinical practice, the treatment sessions were not recorded and there was no associated treatment integrity or competency check. Therefore, the extent to which therapists adhered to the treatment protocol during each treatment is uncertain, and the degree of clinical supervision received for each case is unknown. Some therapists may be reluctant to employ trauma-focussed therapeutic techniques due to their concerns that trauma-focussed therapy may be unsuitable or potentially harmful for some patients with PTSD (Murray et al., 2022).

Some patients also accessed brief low intensity interventions before starting high intensity treatment (65.79% of the Tf-CBT group and 44.74% of the EMDR group). NICE (2018b) guidelines do not currently recommend low intensity interventions for the treatment of PTSD due to limited evidence for their effectiveness in treating this condition (NICE, 2018a). Further, Robinson et al. (2020) found that among 935 patients with PTSD who accessed low intensity CBT at NHS Talking Therapies services, only 4.8% attained a reliable and clinically significant improvement in symptoms. Pre-treatment measures in the current study were taken from all patients' first high intensity treatment session, however, it is possible that accessing low intensity treatment prior to this may have had some effect on patients that did so. For some patients, accessing low intensity treatment could have reduced their PTSD symptoms or otherwise prepared them for therapeutic change. Alternatively, if low intensity treatment was ineffective, this could negatively alter patients' expectations regarding the efficacy of psychological therapies, thereby reducing their likelihood of reliable change following access to high intensity treatment (Constantino et al., 2011). Excluding these patients may have increased internal reliability but would have reduced the EMDR group to an unfeasibly small size and produced a sample that was less representative of NHS Talking Therapies patients.

Nevertheless, the extent to which this sample is representative of the wider population of NHS Talking Therapies patients is unknown, potentially limiting the generalisability of the current findings. However, the difference in the rate of reliable improvement between patients who received their model indicated treatment, and those who did not, was so small (and marginally in favour of patients who received model-indicated suboptimal treatments), that it seems unlikely that testing the PAI in a larger, more representative sample would produce meaningfully different results.

Due to the naturalistic setting, patients were not randomised to treatment. PSM was implemented to control for confounding by indication. But, unlike randomisation, PSM only balances observed covariates. Therefore, it is possible that the two treatment groups systematically differed on unobserved covariates.

3.4.3 Theoretical Considerations

There is debate as to whether Tf-CBT and EMDR act through distinct mechanisms (Landin-Romero et al., 2018). If EMDR and Tf-CBT share the same mechanisms of change this could mean that there is no interaction between patient characteristics and treatment selection, and the finding of Deisenhofer et al. (2018) could be an artefact of overfitting. This is congruent with the *common factors* model, which implies that the factors shared by different forms of trauma-focussed psychological therapy are necessary and sufficient to facilitate therapeutic change in PTSD, and the factors that distinguish different forms of trauma-focussed therapy are relatively insignificant (Wampold, 2019). However, a recent meta-analysis by Nye et al. (2023) found a small but significant superiority of personalised treatment over treatment as usual, and when scaled up to the magnitude of a national level delivery programme such as NHS Talking Therapies, such small differences become significant (Barkham, 2023).

3.4.4 Future Directions

Future studies should use a larger sample with a PTSD symptom measure as outcome, test different modelling methods, apply bootstrapping during model development and internal validation, and then externally cross-validate in either a hold-out test sample, data from another location (i.e., geographic validation), or data collected at a later time (i.e., temporal validation). Once the models have been externally validated in larger samples, an even more rigorous test is the prospective application and validation of such models by assigning incoming patients to the treatment recommended by the model and comparing this data-informed allocation to a random or clinically intuitive decision (Delgadillo et al., 2022; Lutz et al., 2022).

In the current dataset, *n* = 38 patients accessed EMDR as their only high intensity treatment for PTSD, whereas *n* = 273 patients accessed EMDR as part of an integrated treatment with CBT for PTSD. This suggests that EMDR is most often delivered as part of an integrated cognitive behavioural treatment for PTSD. In which case it would be pertinent to investigate differential treatment response to Tf-CBT versus Tf-CBT with integrated EMDR, similar to the way that Hoeboer et al. (2021) investigated differential response to PE alone versus PE plus skills training.

3.4.5 Conclusions

The PAI model developed by Deisenhofer et al. (2018) does not generalise beyond the model development sample. Since the external validation presented in this paper is limited by a small sample size, the findings must be interpreted with caution. Nevertheless, this study highlights the importance of external validation in prediction modelling. Additionally, it emphasizes important factors to consider during model development, such as sample size, prediction method, and validation methods. This study underlines the need for clinicians to routinely administer PTSD specific outcome measures before, during and on completion of trauma treatments in routine practice, so that sufficient data for prediction modelling research is made available. Finally, this study highlights the need for researchers to develop and then externally validate clinical prediction models for those trauma treatments which are most typically delivered in routine services, in order for the models to have maximum applied utility for those delivering and receiving these treatments.

CHAPTER 4
A comparison of machine learning methods at predicting the outcome of trauma-focussed cognitive behavioural therapy

4.1 Introduction

The systematic review presented in Chapter 2 found that only two studies compared the performance of machine learning (ML) methods against that of simpler, traditional statistical methods (Held et al., 2022; Stuke et al., 2021). One of these studies found that five different ML methods performed better than logistic regression, whereas the other found that linear regression performed best but only tested one ML method (see section 2.4.4). This suggests that some ML methods may be better than others at predicting treatment outcomes for post-traumatic stress disorder (PTSD), and some may be no better than linear regression. Given the sample size issues highlighted in the review, the optimal prediction method remains unclear.

Furthermore, neither of the above studies attempted to externally validate their prediction models, as was the case for all but one study in the review. The study presented in Chapter 3 highlights the necessity of external validation of clinical prediction models, as the personalised advantage index (PAI) did not generalise to the external validation sample. Potential reasons for this include [1] the choice of ML method, [2] the training sample size, and [3] inadequate internal cross-validation.

The aim of the current study was to test this by comparing the accuracy of different ML models and linear regression at predicting the outcome of trauma-focussed cognitive behavioural therapy (Tf-CBT) in a geographic validation sample. Adhering to the recommendations in sections 2.4.5 and 3.4.5, a sample size calculation was performed to ensure that the training and validation samples were of adequate size, bootstrapping was applied for internal cross-validation, and a PTSD symptom measure was applied as an outcome measure. To investigate the effect of sample size, all models were trained on samples of iteratively restricted size and tested in the external validation sample. This study was conducted in accordance with the machine learning pipeline explained in Chapters 1 and 2 (Delgadillo & Atzil-Slonim, 2022) and reported following TRIPOD+AI guidelines (Collins et al., 2024). The pre-registered research questions were: [1] Do ML methods predict the outcome of Tf-CBT with greater accuracy than linear regression? [2] Are some ML methods better than others at predicting the outcome of Tf-CBT? [3] Is this difference moderated by sample size?

4.2 Method

4.2.1 Pre-registration

The study research questions, dataset information, and analysis plan were pre-registered with the Open Science Framework (OSF) and the pre-registration can be viewed here: https://osf.io/mc4n6

4.2.2 Participants, Setting and Intervention

The dataset used in this study included anonymised clinical case records of *N* = 2,064 patients who accessed one of sixteen NHS Talking Therapies services across seven NHS Trusts in Cambridgeshire, Cheshire, Lancashire, London, and Yorkshire and Humber, between June 2010 and April 2017. Patients whose data were included in the study sample had a provisional ICD-10 diagnosis of PTSD (WHO, 2019) and completed a measure of PTSD symptoms at initial assessment. To be included in the study sample patients were required to have received ≥ 1 session of Tf-CBT. The total eligible study sample consisted of *N* = 1,319 patients.

4.2.3 Measures

Psychometric measures below were administered at every treatment and assessment session. Pre-treatment scores and predictor information were extracted from each patient's initial assessment session data. Post-treatment outcome scores were extracted from each patient's final treatment session data.

4.2.3.1 Outcome Measures

4.2.3.1.1 Patient Health Questionnaire 9

For comparison with previous studies (Chapter 3; Deisenhofer et al., 2018), prediction models were trained with the Patient Health Questionnaire 9 (PHQ-9; Kroenke et al., 2001) as a proxy indicator of PTSD treatment outcome. The PHQ-9 is a validated, nine-item, self-report measure of depression severity. The PHQ-9 is routinely administered at every NHS Talking Therapies session, and due to its high correlation with PTSD measures it was taken as a proxy outcome measure of PTSD treatment response in prior studies where a measure of PTSD symptoms was not adequately applied in the study dataset (Deisenhofer et al., 2018; Tait et al., 2024).

4.2.3.1.2 Impact of Event Scale-Revised

The primary outcome measure was the Impact of Event Scale-Revised (IES-R; Weiss, 2007). The IES-R is a validated, 22-item, self-report measure of PTSD severity. Total IES-R scores range from 0-88, with higher scores indicating greater number and frequency of PTSD symptoms. Creamer et al. (2003) reported good reliability in a mixed treatment-seeking/community sample (Cronbach's a = 0. 96), and good convergent validity with the PTSD Checklist (r = .84). A change of ≥ 9 points on the IES-R has been recommended as an index of reliable improvement or deterioration in symptoms (NHS England, 2014).

4.2.3.2 Predictors

All demographic information and validated measures that were collected at initial assessment were included as predictors. This included pre-treatment scores on the PHQ-9 and IES-R, and the following variables:

4.2.3.2.1 Generalised Anxiety Disorder 7

The Generalised Anxiety Disorder 7 (GAD-7; Spitzer et al., 2006) is a validated seven-item self-report measure of anxiety symptoms, scores range from 0-21 with higher scores indicating greater number and frequency of anxiety symptoms.

4.2.3.2.2 Work and Social Adjustment Scale

The Work and Social Adjustment Scale (WSAS; Mundt et al., 2002) is a validated five-item self-report measure of the extent to which a person's mental health problems impair their daily functioning. WSAS scores range from 0-40 with higher scores indicating greater impairment across more domains of daily functioning.

4.2.3.2.3 Demographic and Health Variables

Age, gender, ethnicity, disability, long-term condition (LTC), employment, and antidepressant medication data were extracted from patient records. Ethnicity was based on the Office for National Statistics ethnic categories (Office for National Statistics, n.d.) and was self-reported. Disability was a binary indicator of the presence of any patient-reported disability. LTC was an indicator of whether a patient had a long-term physical health condition such as diabetes, arthritis, or a chronic respiratory condition.

Employment status was reduced to a binary indicator of unemployment, with “Unemployed” and “Long-term sick” in one category, and “Employed”, “Student”, “Homemaker/carer”, “Voluntary work”, and “Retired” in the other. Medication status was reduced to a binary indicator of taking medication, with “prescribed and taking” in one category, and “prescribed but not taking” and “not prescribed” in the other. Ethnicity was reduced to a binary variable with “White” in one category, and “Asian”, “Black”, “Mixed/Multiple”, and “Other ethnic group” in the other category. Disability and LTC were reduced to binary indicators of disability and LTC respectively.

Local Layer Super Output Area was extracted from patient records and converted to Indices of Multiple Deprivation (IMD; Department for Communities and Local Government, 2015) decile. The IMD ranks geographic areas by their level of social deprivation across multiple domains (e.g., income, education, health, crime), with the 1st decile indicating the 10% most socially deprived areas in England, and the 10th decile indicating the 10% least deprived areas. Nine polynomial functions (one less than the number of levels) were fit for IMD (x^1, x^2, x^3 … x^9) to allow models to capture non-linear relationships between deprivation and outcome.

4.2.4 Sample Size Calculations

The minimum required training sample size was estimated using the R package pmsampsize, which applies the method outlined by Riley et al. (2020). The number of candidate predictor parameters was 19, corresponding to five continuous variables (age, GAD-7, IES-R, PHQ-9, WSAS), five binary variables (unemployment, ethnicity, gender, LTC, medication), and the nine polynomial functions for IMD. Estimates of R^2, intercept, and standard deviation for the sample size calculation were taken from the Tf-CBT model developed by Deisenhofer et al. (2018) in the Tf-CBT group used for that study (adjusted R^2 = .36; final session PHQ-9 mean = 9.42, SD = 7.59). Recommended default values were applied for the remaining parameters. This indicated a minimum required training sample N = 337. For the validation sample, Collins et al. (2016) recommend a minimum sample N = 100, ideally *N* = 200, and Steyerberg (2019) recommends a minimum *N* = 250. Therefore, a minimum validation sample *N* = 250 was sought.

4.2.5 Partitioning the Data into Training and Validation Datasets

The data was split by NHS Trust into a training sample, for developing prediction models, and a validation sample, for testing the models' out-of-sample performance. Splitting the data by NHS Trust allows for *geographic validation* (Steyerberg, 2019). This is a more rigorous test of generalisability than randomly splitting the data, as the training and validation samples will include patients treated by different therapists, at different NHS Talking Therapies services, in different geographic locations.

To select NHS Trusts for the validation sample, Trusts were numbered and numbers were drawn at random using the `sample()` function in the base package in R. However, given prior evidence that pre-treatment PTSD severity and social deprivation are associated with PTSD treatment outcomes, Trusts were drawn at random until a validation sample was identified that exceeded the minimum recommended sample size (*N* = 250) and was not significantly different to the training sample in pre-treatment IES-R score or IMD decile. The validation sample included data from two Trusts with a combined *N* = 464; and the training sample included data from the remaining five Trusts with a combined *N* = 855 (corresponding to an outcome event per variable ratio of 855:19 = 45:1).

4.2.6 Missing Data and Multiple Imputation

The proportion of missing values on each variable in the training and validation samples is tabulated in Appendix J. Disability data was missing for 78.36% of patients in the training dataset and 89.87% of patients in the validation dataset. For this reason, Disability was omitted from all further analyses. Three predictor variables had > 5% missing values in the training sample, these were LTC (33.10%), Medication (12.63%), and Employment

(7.95%). The only predictor with > 5% missing values in the validation sample was LTC (12.50%). Post-treatment IES-R score was missing for 50.53% of the training dataset and 43.53% of the validation dataset. Little's Missing Completely At Random (MCAR) test (performed using the naniar package) indicated that missing values were not MCAR in the training sample (x^2 (3228) = 3960, $p < .001$) or validation sample (x^2 (1162) = 1526, $p < .001$).

Multiple imputation of missing values was performed using the missForest package in R (Stekhoven & Bühlmann, 2012) with default hyperparameter settings applied. To prevent data leakage, multiple imputation was performed separately for the training and validation samples. The imputation models included all predictors and outcome measures; post-treatment scores for GAD-7, WSAS, Employment, and Medication; PHQ-9, GAD-7, and WSAS scores from first and final Tf-CBT sessions; and the number of Tf-CBT sessions attended. The Spearman's correlations between variables before imputation, tabulated in Appendix K, indicate that there were moderate to strong correlations (ρ = .7 to .8) between post-treatment IES-R score and post-treatment scores on GAD-7, PHQ-9, and WSAS. Out-of-bag error estimates for the training sample were NRMSE = .31 and PFC = .32, and for the validation sample NRMSE = .30 and PFC = .32.

4.2.7 Sample Characteristics after Multiple Imputation

Table 4.1 presents sample characteristics for the training and validation samples after imputation of missing values. Continuous variables were not normally distributed (see Appendix L for details), therefore differences between samples were tested using Wilcoxon Rank-Sum tests for continuous and ordinal variables, and chi-square tests for binary variables. The validation sample was significantly different to the training sample on four variables: pre-treatment WSAS score (W = 173,969, p < .001), ethnicity (X^2 (1) = 30.99, p

< .001), and the proportion taking medication at both pre-treatment (X^2 (1) = 24.82, p < .001) and post-treatment (X^2 (1) = 31.44, p < .001). This indicates that the validation sample reported greater functional impairment, a smaller majority of the validation sample identified as white, and a smaller proportion of the validation sample reported taking anti-depressant medication. These differences reflect natural variation that would be expected between services and are appropriate for testing the validity and transportability of prediction models.

Table 4.1

Sample Characteristics After Multiple Imputation

Variable	Training sample (*N* = 855)	Validation sample (*N* = 464)
	Median (IQR)	Median (IQR)
Age	37.00 (27.00 - 49.00)	37.50 (28.00 - 49.00)
IES-R (pre-treatment)	63.00 (50.00 - 72.50)	64.00 (51.00 - 72.00)
IES-R (post-treatment)	37.00 (18.00 - 58.00)	33.00 (15.00 - 61.00)
PHQ-9 (pre-treatment)	18.00 (13.50 - 22.00)	18.00 (13.75 - 22.00)
PHQ-9 (post-treatment)	10.00 (4.00 - 18.00)	9.00 (4.00 - 16.00)
GAD-7 (pre-treatment)	17.00 (13.00 - 19.00)	16.00 (13.00 - 19.00)
GAD-7 (post-treatment)	9.00 (4.00 - 16.00)	8.50 (4.00 - 16.00)
WSAS (pre-treatment)	23.00 (15.00 - 30.00)	26.00 (17.00 - 32.00)*
WSAS (post-treatment)	15.00 (6.00 - 25.00)	15.00 (6.00 - 25.25)
Tf-CBT sessions attended	9.00 (5.00 - 13.00)	10.00 (5.00 - 14.00)
IMD decile	3.00 (2.00 - 6.00)	4.00 (2.00 - 6.00)
	N (%)	*N* (%)
Gender (female)	519 (60.70%)	277 (59.70%)
Ethnicity (white)	723 (84.56%)	332 (71.55%)*
Unemployed (pre-treatment)	314 (36.73%)	158 (34.05%)

Unemployed (post-treatment)	302 (35.32%)	167 (35.99%)
Taking Medication (pre-treatment)	478 (55.91%)	192 (41.38%)*
Taking Medication (post-treatment)	506 (59.18%)	199 (42.89%)*
Long Term Condition	254 (29.71%)	154 (33.19%)
IMD decile 1	157 (18.36%)	22 (4.74%)
IMD decile 2	153 (17.89%)	106 (22.84%)
IMD decile 3	118 (13.80%)	96 (20.69%)
IMD decile 4	94 (10.99%)	56 (12.07%)
IMD decile 5	67 (7.84%)	59 (12.72%)
IMD decile 6	75 (8.77%)	39 (8.41%)
IMD decile 7	37 (4.33%)	14 (3.02%)
IMD decile 8	64 (7.49%)	28 (6.03%)
IMD decile 9	45 (5.26%)	25 (5.39%)
IMD decile 10	45 (5.26%)	19 (4.09%)

Note. GAD-7 = Generalised Anxiety Disorder 7; IES-R = Impact of Event Scale – Revised; IMD = Index of Multiple Deprivation; IQR = Inter-quartile Range; PHQ-9 = Patient Health Questionnaire 9; Tf-CBT = Trauma-focussed Cognitive Behavioural Therapy; WSAS = Work and Social Adjustment Scale.

Post-treatment measures were included in the multiple imputation models but not as predictors in the outcome prediction models.

* Significantly different to the training sample, $p < .001$ (tested using Wilcoxon Rank Sum Test for continuous/ordinal variables and Chi-square Test for categorical variables)

4.2.8 Data analysis strategy

All analyses were performed using *RStudio* with R version 4.4.1.

4.2.8.1 Training Prediction Models

Eight ML algorithms were selected to represent different families of supervised ML: Elastic Net (EN; *Penalised regressions*), Random Forest (RF; *Decision trees*), Boosted Generalised Linear Model (BoostGLM; *Boosted models*), Bayesian Generalised Linear Model (BayesGLM; *Bayesian models*), Radial Support Vector Machine (RSVM; *Linear models*), Multi-Layer Perceptron (MLP; *Deep learning*), Bayesian Regularised Neural Network (BRNN; *Deep learning/Bayesian models*), Genetic Regression (GR; *Evolutionary algorithms*). GR was included for comparison with the model developed by Deisenhofer et al. (2018) and externally validated in Chapter 3. The remaining methods were chosen because previous studies found that they performed comparatively well at predicting psychological therapy outcomes (Bennemann et al., 2022; Giesemann et al., 2023; Gómez Penedo et al., 2023; Held et al., 2022; Webb et al., 2020). An additional deep learning/Bayesian model was included (BRNN) as the systematic review in Chapter 2 found that no previous studies tested Bayesian or deep learning models. Linear Regression (LR) was included as a non-ML comparator to explore the extent to which ML algorithms improve prediction model performance in this context. See Appendix M for further details of the prediction methods.

Models were trained and tested using the caret package. The internal cross-validation method was bootstrap .632 (Efron, 1983) with 1000 repetitions, and hyperparameter selection was optimised via grid search with the default grid settings for each method and the default evaluation metric (Root Mean Square Error). For the GR model, predictors were selected using the genetic algorithm option in the glmulti package (with confsetsize = 512 and default settings for all other parameters). Predictors with importance > .8 were then entered into a linear regression model (lm), which was trained and internally cross-validated using the caret package in the same way as the other models. This was done to replicate the methods used by Deisenhofer et al. (2018) when developing the model tested in Chapter 3, while retaining the same cross-validation procedure for comparability with other models in this study. This facilitated comparison of cross-validation procedure (bootstrap .632 vs. leave-one-out) and exploration of the effect of training sample size on GR model performance.

4.2.8.2 Evaluating Model Performance

Prediction model performance was evaluated via R^2 and Root Mean Square Error (RMSE). R^2 is a measure of average prediction accuracy, which ranges from 0 to 1 with higher values indicating greater accuracy. R^2 can also be interpreted as the proportion of variance in the outcome that is explained by the model. RMSE is a measure of average prediction error, which ranges from 0 to infinity with higher values indicating greater error. RMSE is the square root of the mean of the squared differences between the observed and predicted values and is relative to the scale of the outcome measure.

For each method the hyperparameter settings with the lowest RMSE in internal-cross validation were selected as the final model. Final models were applied to predict outcomes in the validation sample using the `predict()` function in the stats package, and R^2 and RMSE were calculated using the `postResample()` function in the caret package. As the aim was to identify the model that makes the most accurate predictions in new data, model performance was evaluated in the validation sample. Internally cross-validated performance within the training sample was also compared to examine *shrinkage* (the extent to which the models' performance estimates diminish between the training and validation sample). Model *calibration* (the level of agreement between predicted and observed outcomes in the validation sample) was compared via calibration plots.

4.2.8.3 Evaluating Predictor Variable Importance

Where possible, normalised variable importance values were extracted using the `varImp()` function in *caret*, to explore the relative importance of each of the predictor variables to the respective model. Model specific variable importance metrics were not available for MLP, BRNN, RSVM, or BayesGLM. Additionally, model coefficients and related statistics were extracted for the linear models (LR, GR, BayesGLM, BoostGLM, EN) to examine the direction of the relationships between important predictors and outcome.

4.2.8.4 Exploring the Effect of Training Sample Size

To examine the effect of training sample size on model performance the training sample was iteratively restricted by nested random sampling. Models were trained on each of the restricted training samples, applying internal cross-validation and hyperparameter optimisation procedures each time, and evaluated in the full validation sample (*N* = 464). None of the randomly selected iteratively restricted training samples were significantly different to the validation sample on pre-treatment IES-R score of IMD decile.

4.2.8.5 Outliers

To maximise the available sample size and to evaluate the calibration of prediction models in a way that reflects data distribution trends in routine clinical care, outliers were not removed or modified in the primary analysis. As pre-registered, outliers on the outcome variables (IES-R and PHQ-9) were sought by inspecting robust (median absolute deviation based) *z* scores for any exceeding the threshold of 3.29 (Thériault et al., 2024). By this method, no outliers were identified in the training or test sample, therefore sensitivity analysis with Winsorised values was not necessary.

4.2.9 Ethical Approval

Ethical approval was granted by the London - City & East Research Ethics Committee (Ref: 15/LO/2200) for this data to be used for research. All patients in the dataset provided documented verbal consent for their anonymised data to be used for research.

4.3 Results

4.3.1 Participants and Treatment Outcomes

The total analysed sample consisted of $N = 1,319$ patients with PTSD who had received ≥ 1 session of Tf-CBT; $n = 855$ in the training sample and $n = 464$ in the validation sample. Sample selection is depicted in the STROBE diagram in Figure 4.1. The median number of Tf-CBT sessions in the training sample was 9 (interquartile range = 5 – 13) and the mode was 11 (range = 1 – 49). The median number of Tf-CBT sessions in the validation sample was 10 (interquartile range = 5 – 14) and the modes were 13 and 14 (range = 1 – 31). Pre- to post-treatment effect sizes are presented in Table 4.2. The overlap of the 95% confidence intervals suggests no significant difference in treatment effect size on either the IES-R or PHQ-9 between the training and validation sample.

Table 4.2

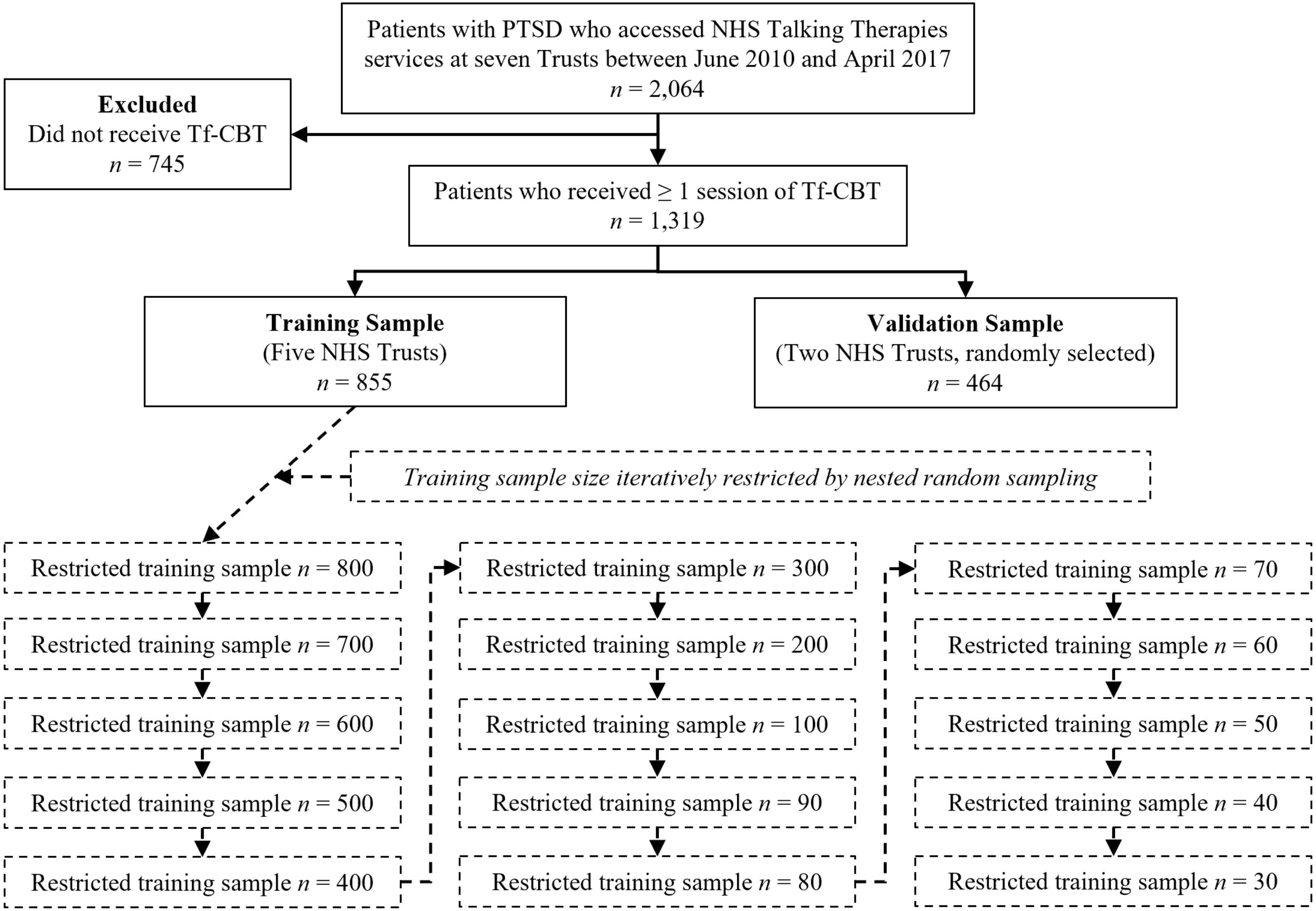
Comparison of Treatment Effect Sizes

Measure	Sample	*d*	95% Confidence Interval	
			Lower limit	Upper limit
PHQ-9	Training (N = 855)	1.07	0.98	1.16
	Validation (N = 464)	1.19	1.06	1.31
IES-R	Training (N = 855)	1.32	1.23	1.41
	Validation (N = 464)	1.30	1.17	1.43

Note. IES-R = Impact of Event Scale – Revised; PHQ-9 = Patient Health Questionnaire 9.
Effect sizes and confidence intervals calculated using the method described by Minami et al. (2008) with Spearman’s correlations.

Figure 4.1

STROBE Flow Diagram Depicting the Sample Selection Process


Patients with PTSD who accessed NHS Talking Therapies services at seven Trusts between June 2010 and April 2017
n = 2,064
Excluded
Did not receive Tf-CBT
n = 745
Patients who received ≥ 1 session of Tf-CBT
n = 1,319
Training Sample
(Five NHS Trusts)
n = 855
Validation Sample
(Two NHS Trusts, randomly selected)
n = 464
Training sample size iteratively restricted by nested random sampling
Restricted training sample n = 800
Restricted training sample n = 700
Restricted training sample n = 600
Restricted training sample n = 500
Restricted training sample n = 400
Restricted training sample n = 300
Restricted training sample n = 200
Restricted training sample n = 100
Restricted training sample n = 90
Restricted training sample n = 80
Restricted training sample n = 70
Restricted training sample n = 60
Restricted training sample n = 50
Restricted training sample n = 40
Restricted training sample n = 30

4.3.2 Evaluating Model Performance

Table 4.3 presents the performance evaluation metrics for each model trained on the full training sample (N = 855), predicting final session PHQ-9 score and final session IES-R score. Internal metrics refer to the internally cross-validated model performance within the training sample, external metrics refer to model performance in the validation sample, and shrinkage refers to the decrease in the performance estimate from the training sample to the validation sample. The hyperparameters column presents the optimal hyperparameter settings selected by grid search and applied in the final model.

On both outcome measures, there were only marginal differences in model performance between all models except for RF, which appeared to perform significantly better than the other models in internal cross-validation, but performed worst in external validation, and exhibited the greatest shrinkage. MLP exhibited the least shrinkage on both metrics for both outcome measures and performed best at predicting post-treatment PHQ-9 score in the validation sample (followed closely by BRNN). When predicting post-treatment IES-R score in the validation sample, GR performed best in both RMSE and R^2, and BoostGLM and EN also performed comparably well at predicting both outcomes in the validation sample. When predicting either outcome in the validation sample, LR outperformed RF and RSVM and performed equivalently to BayesGLM on all metrics.

Table 4.3

Comparison of Prediction Model Performance Metrics

Outcome	Model	Hyperparameter settings	RMSE			R^2		
			Internal	External	Shrinkage	Internal	External	Shrinkage
PHQ-9	LR		6.569	6.810	0.241	.292	.220	-.073
	BayesGLM		6.569	6.809	0.241	.292	.220	-.073
	BoostGLM	mstop = 150, prune = no	6.543	6.783	0.240	.297	.224	-.073
	BRNN	neurons = 1	6.544	6.774	0.229	.297	.227	-.071
	EN	alpha = 0.1, lambda = 0.66150307036477	6.548	6.783	0.234	.296	.222	-.074
	GR		6.529	6.796	0.267	.300	.223	-.077
	MLP	hidden1 = 1, n.ensemble = 1	*6.599*	**6.774**	**0.175**	*.288*	**.228**	**-.060**
	RF	mtry = 19	**5.485**	*6.870*	*1.385*	**.471**	*.207*	*-.264*
	RSVM	sigma = 0.0307501295915785, C = 0.25	6.579	6.846	0.268	.296	.216	-.080

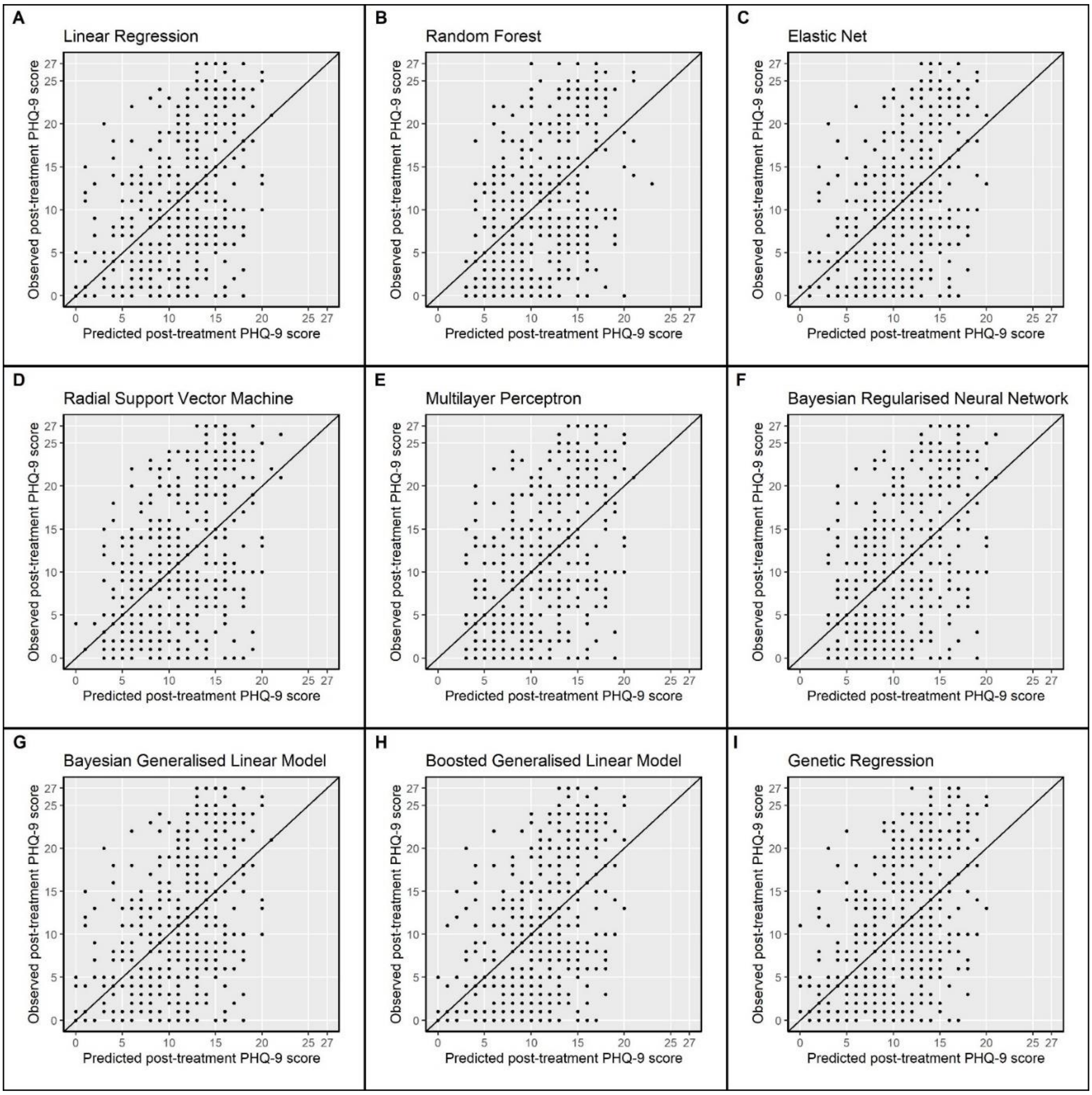
IES-R	LR		19.111	21.780	2.669	.284	.191	-.093
	BayesGLM		19.111	21.780	2.669	.284	.191	-.093
	BoostGLM	mstop = 100, prune = no	19.025	21.606	2.582	.290	.199	-.090
	BRNN	neurons = 1	19.043	21.696	2.653	.289	.195	-.094
	EN	alpha = 0.1, lambda = 2.06631627939549	19.045	21.625	2.579	.288	.197	-.091
	GR		18.983	**21.562**	2.579	.292	**.206**	-.087
	MLP	hidden1 = 1, n.ensemble = 1	*19.264*	21.712	**2.448**	*.277*	.196	**-.081**
	RF	mtry = 19	**15.749**	*22.094*	*6.345*	**.477**	*.171*	*-.305*
	RSVM	sigma = 0.0307501295915785, C = 0.25	19.135	21.842	2.707	.289	.190	-.098

Note. The best performing model on each metric is highlighted in **bold** and the worst is highlighted in *italic*.
Shrinkage is the difference in model performance between internal cross-validation and external validation; performance is typically overestimated in internal cross-validation and this is indicated by an increase in error (RMSE) and decrease in accuracy (R^2) in external validation.
Values are given to three decimal places to allow comparison of small differences in model performance.
IES-R = Impact of Event Scale – Revised; PHQ-9 = Patient Health Questionnaire 9; RMSE = Root Mean Squared Error.
Model: LR = Linear Regression; BayesGLM = Bayesian Generalised Linear Model; BoostGLM = Boosted Generalised Linear Model; BRNN = Bayesian Regularised Neural Network; EN = Elastic Net; GR = Genetic Regression; MLP = Multi-Layer Perceptron; RF = Random Forest; RSVM = Radial basis function Support Vector Machine.

Figures 4.2 and 4.3 present the calibration plots for models predicting PHQ-9 and IES-R outcomes in the validation sample. These plots show that although the observed outcomes for both measures span the full range of the respective scale, none of the models accurately predict scores at the higher end of the severity scale. Further, the non-linear models (RF, MLP, BRNN, RSVM) do not accurately predict scores at the extreme low end of the severity scale (with the exception of the RSVM predicting PHQ-9). Predicted scores appear to be clustered around the centre of the distribution of the outcome measures in the training sample (see Table 4.1).

Figure 4.2

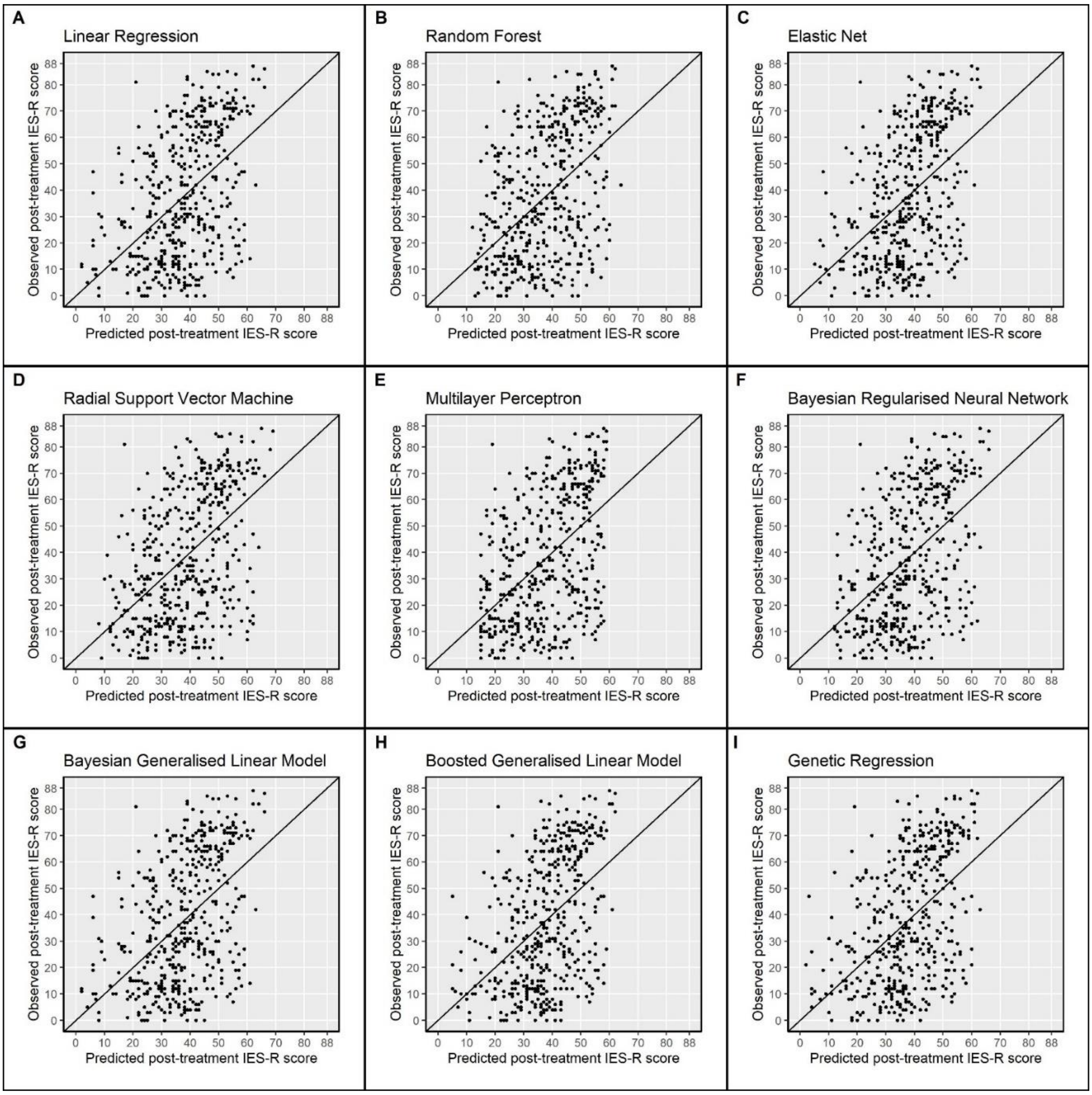
Calibration Plots for Models Predicting Post-Treatment PHQ-9 Score in the Validation Sample



Note. PHQ-9 = Patient Health Questionnaire 9.

Figure 4.3

Calibration Plots for Models Predicting Post-Treatment IES-R Score in the Validation Sample



Note. IES-R = Impact of Event Scale – Revised.

4.3.4 Evaluating Predictor Variable Importance

Available normalised variable importance values for models predicting final session PHQ-9 and IES-R are presented in Table 4.4. LR, GR, and RF found that pre-treatment IES-R score was the most important predictor of treatment outcome, whether outcome was measured with the PHQ-9 or IES-R. However, BoostGLM and EN found that unemployment was the most important predictor of outcomes, and pre-treatment IES-R score was less important than medication, LTC, IMD and gender. Unemployment was the second most important predictor in the LR and GR models. RF attributed greater importance to GAD-7 and WSAS than the linear models did.

Coefficients for the linear models predicting final session PHQ-9 and IES-R score are presented in Tables 4.5 and 4.6 respectively. GR selected pre-treatment IES-R score, unemployment, pre-treatment PHQ-9 score, and age as predictors of post-treatment IES-R score, and these same four predictors plus medication as predictors of post-treatment PHQ-9 score. Inspection of the coefficients reveals that there was very little difference between the BayesGLM and LR models, and the coefficients for the variables selected by GR were similar across the five linear models. There was some agreement and some discrepancy in the variables that were not included by BoostGLM, and coefficients shrunk to zero by EN.

Across the linear models, the coefficients suggest that higher pre-treatment IES-R score, being unemployed, higher pre-treatment PHQ-9 score, and younger age, may be associated with higher scores on both the PHQ-9 and IES-R at post-treatment. Taking anti-depressant medication appears to be associated with higher post-treatment PHQ-9 score, but not IES-R score. There is some indication that social deprivation may be associated with final session PHQ-9 score, and to a lesser extent with final session IES-R score, and that the seventh order polynomial term best fits this relationship.

Table 4.4

Model Specific Normalised Variable Importance Values for Models Predicting Post-treatment PHQ-9 Score and Post-treatment IES-R Score

	PHQ-9					IES-R				
	LR	GR	BoostGLM	EN	RF	LR	GR	BoostGLM	EN	RF
IES-R	100	100	4.39	4.34	100	100	100	4.59	4.55	100
Unemployed	64.47	64.08	100	100	24.12	61.56	58.29	100	100	22.33
PHQ-9	50.50	71.86	9.60	8.84	68.25	20.75	17.89	4.37	4.10	30.60
GAD7	0.14			1.18	28.23	0.36		0.00	0.88	19.56
WSAS	7.36		0.50	0.93	42.37	4.85		0.34	0.64	31.50
Medication	36.63	7.92	51.10	54.18	5.60	15.22		18.54	24.22	1.21
Age	36.15	0.00	1.51	1.56	45.69	34.33	0.00	1.37	1.62	36.38
LTC	11.17		6.66	10.84	2.28	17.85		16.16	23.80	1.44
Gender	24.62		25.57	29.73	3.18	12.82		7.56	15.51	1.52
Ethnicity	6.79			4.25	0.00	0.21		0.00	0.00	0.00
IMD decile (^1)	16.51		33.33	39.53	2.39	1.93		0.00	8.42	2.33
IMD decile (^2)	10.69		7.21	14.80	2.19	9.45		1.72	14.60	0.42
IMD decile (^3)	17.31		22.33	27.91	2.27	15.56		16.22	27.74	2.14
IMD decile (^4)	8.39		6.78	12.90	2.90	4.53		1.89	10.97	0.78
IMD decile (^5)	7.96			3.70	2.70	0.00		-	0.00	0.84
IMD decile (^6)	0.00			0.00	0.92	5.26		-	6.63	0.30
IMD decile (^7)	25.32		40.98	44.02	3.19	16.54		16.91	27.50	2.26
IMD decile (^8)	4.52			0.00	2.05	6.22		-	2.08	0.43
IMD decile (^9)	8.99		4.66	11.86	4.50	5.49		-	5.88	3.32

Note. GAD-7 = Generalised Anxiety Disorder 7; IES-R = Impact of Event Scale – Revised; IMD = Index of Multiple Deprivation; LTC = Long Term Condition; PHQ-9 = Patient Health Questionnaire 9; WSAS = Work and Social Adjustment Scale.
Models: LR = linear regression; GR = genetic regression; BoostGLM = boosted generalised linear model; EN = elastic net; RF = random forest.

Table 4.5

Coefficients for Linear Models Predicting Final Session PHQ-9 Score

	LR				GR				BayesGLM				BoostGLM	EN
	B	*SE*	*t*	*p*	*B*	*SE*	*t*	*p*	*B*	*SE*	*t*	*p*	*B*	*B*
Intercept	-0.77	1.24	-0.62	.532	-1.88	1.10	-1.71	.087	-0.78	1.24	-0.63	.532	-12.20	-0.12
IES-R	0.12	0.02	8.22	<.001	0.13	0.01	8.60	<.001	0.12	0.02	8.22	<.001	0.12	0.10
Unemployed	2.75	0.52	5.32	<.001	3.18	0.49	6.52	<.001	2.75	0.52	5.32	<.001	2.76	2.34
PHQ-9	0.26	0.06	4.18	<.001	0.30	0.04	6.97	<.001	0.26	0.06	4.18	<.001	0.27	0.20
Age	-0.05	0.02	-3.01	.003	-0.05	0.02	-2.81	.005	-0.05	0.02	-3.01	.003	-0.04	-0.02
Medication (taking)	1.49	0.49	3.05	.002	1.56	0.48	3.26	.001	1.49	0.49	3.05	.002	1.41	1.22
Gender (female)	-0.97	0.47	-2.07	.039					-0.97	0.47	-2.07	.039	-0.71	-0.49
LTC	0.51	0.53	0.97	.333					0.51	0.53	0.97	.333	0.18	0.00
GAD-7	0.00	0.07	0.07	.946					0.00	0.07	0.07	.946		0.06
WSAS	0.02	0.03	0.66	.511					0.02	0.03	0.66	.511	0.01	0.03
Ethnicity (other)	-0.39	0.64	-0.61	.542					-0.39	0.64	-0.61	.542		0.00
IMD (^1)	-1.17	0.83	-1.40	.161					-1.17	0.83	-1.40	.161	-0.92	-0.87
IMD (^2)	0.75	0.81	0.93	.353					0.75	0.81	0.93	.353	0.20	0.00
IMD (^3)	1.16	0.79	1.47	.142					1.16	0.79	1.47	.142	0.62	0.19
IMD (^4)	-0.57	0.77	-0.74	.459					-0.57	0.77	-0.74	.459	-0.19	-0.12
IMD (^5)	-0.58	0.82	-0.71	.480					-0.58	0.82	-0.71	.481		0.00

IMD (^6)	0.04	0.78	0.06	.955					0.04	0.78	0.06	.955		0.00
IMD (^7)	1.62	0.76	2.12	.034					1.62	0.76	2.12	.034	1.13	0.63
IND (^8)	0.35	0.83	0.43	.671					0.35	0.83	0.42	.672		0.00
IMD (^9)	0.64	0.81	0.79	.430					0.64	0.81	0.79	.430	0.13	0.00

Note. *B* = unstandardised coefficient; *SE* = Standard Error.
GAD-7 = Generalised Anxiety Disorder 7; IES-R = Impact of Event Scale – Revised; IMD = Index of Multiple Deprivation; LTC = Long Term Condition; PHQ-9 = Patient Health Questionnaire 9; WSAS = Work and Social Adjustment Scale.
Models: LR = linear regression; GR = genetic regression; BoostGLM = boosted generalised linear model; BoostGLM = boosted generalised linear model; EN = elastic net.

Table 4.6

Coefficients for Linear Models Predicting Final Session IES-R Score

	LR				GR				BayesGLM				BoostGLM	EN
	B	*SE*	*t*	*p*	*B*	*SE*	*t*	*p*	*B*	*SE*	*t*	*p*	*B*	*B*
Intercept	4.39	3.61	1.22	.224	1.75	3.19	0.55	.584	4.39	3.61	1.22	.224	-34.09	5.18
IES-R	0.45	0.04	10.35	<.001	0.47	0.04	11.07	<.001	0.45	0.04	10.35	<.001	0.44	0.41
Unemployed	9.62	1.50	6.41	<.001	10.91	1.41	7.74	<.001	9.62	1.50	6.41	<.001	9.62	9.07
PHQ-9	0.40	0.18	2.22	.026	0.56	0.12	4.52	<.001	0.40	0.18	2.22	.026	0.42	0.37
Age	-0.19	0.05	-3.62	<.001	-0.15	0.05	-3.09	.002	-0.19	0.05	-3.62	<.001	-0.13	-0.15
Medication (taking)	2.35	1.42	1.66	.098					2.35	1.42	1.66	.098	1.78	2.20
Gender (female)	-1.93	1.36	-1.41	.158					-1.92	1.36	-1.41	.158	-0.73	-1.41
LTC	2.97	1.54	1.93	.054					2.97	1.54	1.93	.054	1.55	2.16
GAD-7	0.03	0.21	0.13	.893					0.03	0.21	0.13	.893		0.08
WSAS	0.05	0.09	0.59	.552					0.05	0.09	0.59	.552	0.03	0.06
Ethnicity (other)	0.22	1.85	0.12	.905					0.22	1.85	0.12	.905		0.00

IMD (^1)	-0.71	2.42	-0.30	.768		-0.72	2.42	-0.30	.768		-0.76
IMD (^2)	2.50	2.35	1.07	.287		2.50	2.35	1.07	.287	0.17	1.32
IMD (^3)	3.90	2.30	1.69	.091		3.89	2.30	1.69	.091	1.56	2.52
IMD (^4)	-1.26	2.24	-0.56	.574		-1.26	2.24	-0.56	.574	-0.18	-0.99
IMD (^5)	-0.23	2.39	-0.10	.922		-0.23	2.39	-0.10	.923		0.00
IMD (^6)	1.44	2.26	0.64	.524		1.44	2.26	0.64	.525		0.60
IMD (^7)	3.98	2.22	1.79	.073		3.98	2.22	1.79	.074	1.63	2.49
IND (^8)	1.78	2.43	0.73	.463		1.78	2.43	0.73	.463		0.19
IMD (^9)	1.56	2.37	0.66	.510		1.56	2.37	0.66	.510		0.53

Note. *B* = unstandardised coefficient; *SE* = Standard Error.

GAD-7 = Generalised Anxiety Disorder 7; IES-R = Impact of Event Scale – Revised; IMD = Index of Multiple Deprivation; LTC = Long Term Condition; PHQ-9 = Patient Health Questionnaire 9; WSAS = Work and Social Adjustment Scale.

Models: LR = linear regression; GR = genetic regression; BoostGLM = boosted generalised linear model; BoostGLM = boosted generalised linear model; EN = elastic net.

4.3.5 Exploring the Effect of Training Sample Size

Table 4.7 presents prediction error (RMSE), and Table 4.8 presents prediction accuracy (R^2) in the validation sample for models trained on samples of incrementally restricted size. To aid interpretation, this information is also presented as line graphs in Figure 4.4. Model performance appears to interact with training sample size in several ways. RF is among the best performing models when the training sample is small, but then becomes the worst performing model with larger training samples (around $N \geq 600$). Conversely, MLP is consistently the worst performing model with smaller training sample sizes ($N \leq 300$), but then becomes the best model at predicting post-treatment PHQ-9 score once the training sample exceeds $N = 800$. When predicting post-treatment IES-R, GR had the smallest prediction error for training samples $N \geq 300$, and greatest prediction accuracy for training samples $N \geq 80$, except for instances where EN performed better on one metric or the other. BoostGLM and EN appeared to show the least variation in performance across sample sizes and were consistently among the best models.

Table 4.7

Model Prediction Error (RMSE) in the Validation Sample (N = 464) for Models Trained With Incrementally Restricted Training Sample Size

Outcome	Model	Training sample size															
		30	40	50	60	70	80	90	100	200	300	400	500	600	700	800	855
PHQ-9	LR	10.82	9.13	8.58	8.74	8.64	8.11	7.98	7.53	7.20	7.13	6.94	6.87	6.85	6.83	6.81	6.81
	BayesGLM	10.24	9.02	8.54	8.71	8.61	8.09	7.97	7.52	7.20	7.13	6.94	6.87	6.85	6.83	6.81	6.81
	BoostGLM	7.32	7.46	7.38	**7.39**	7.60	7.34	**7.32**	7.07	6.96	**7.02**	6.89	**6.81**	**6.78**	**6.78**	**6.77**	6.78
	BRNN	8.24	8.13	8.27	8.62	7.98	7.82	7.85	7.47	7.17	7.11	6.92	6.85	6.83	6.79	6.78	6.77
	EN	7.71	7.45	7.47	7.43	**7.36**	**7.19**	7.36	**7.07**	**6.91**	7.03	**6.89**	6.82	6.81	6.79	6.78	6.78
	GR	8.43	7.34	7.85	7.74	7.66	7.66	7.65	7.62	7.17	7.02	6.90	6.83	6.82	6.80	6.79	6.80
	MLP	*12.97*	*10.47*	*11.04*	*10.26*	*10.12*	*9.21*	*9.46*	*8.46*	*7.56*	*7.16*	6.95	6.88	6.85	6.81	6.78	**6.77**
	RF	**7.25**	**7.27**	**7.35**	7.44	7.45	7.47	7.50	7.26	6.96	7.04	*6.99*	6.88	6.83	6.84	*6.85*	*6.87*
	RSVM	7.68	7.51	7.41	7.59	7.52	7.61	7.59	7.39	7.05	7.08	6.94	*6.89*	*6.90*	*6.84*	6.83	6.85
IES-R	LR	29.07	26.96	25.38	24.98	25.19	24.44	24.15	23.07	22.75	22.63	22.07	22.03	21.87	21.80	21.88	21.78
	BayesGLM	27.99	26.64	25.25	24.92	25.14	24.40	24.13	23.06	22.75	22.62	22.07	22.03	21.87	21.80	21.88	21.78
	BoostGLM	**22.48**	23.25	23.10	22.69	23.30	22.67	**22.34**	22.19	21.87	22.02	21.80	21.83	21.70	21.64	21.72	21.61
	BRNN	23.59	23.87	24.14	24.53	24.69	23.84	23.92	22.92	22.57	22.48	21.98	21.93	21.78	21.71	21.80	21.70
	EN	22.64	23.46	22.95	22.63	22.72	22.53	22.62	**22.16**	**21.71**	22.21	21.82	**21.80**	21.69	**21.62**	**21.69**	21.62
	GR	22.91	22.77	22.91	22.88	24.09	22.46	22.51	22.29	21.54	**21.71**	**21.64**	21.94	**21.61**	21.62	21.80	**21.56**
	MLP	*32.45*	*34.20*	*33.40*	*34.07*	*33.59*	*26.72*	*25.86*	*24.54*	*23.40*	*23.42*	22.06	21.99	21.84	21.77	21.83	21.71
	RF	22.99	**22.18**	**22.39**	22.36	22.54	**22.87**	23.28	22.68	22.12	22.07	21.88	21.99	*22.01*	*22.11*	*22.09*	*22.09*
	RSVM	22.81	22.64	22.87	**23.20**	**23.12**	23.43	23.67	22.92	23.01	23.00	*22.19*	*22.25*	21.99	21.90	21.92	21.84

Note. The best performing model in each training sample is highlighted in **bold** and the worst is highlighted in *italic*.

IES-R = Impact of Event Scale – Revised; PHQ-9 = Patient Health Questionnaire 9; RMSE = Root Mean Squared Error.

Model: LR = Linear Regression; BayesGLM = Bayesian Generalised Linear Model; BoostGLM = Boosted Generalised Linear Model; BRNN = Bayesian Regularised Neural Network; EN = Elastic Net; GR = Genetic Regression; MLP = Multi-Layer Perceptron; RF = Random Forest; RSVM = Radial basis function Support Vector Machine.

Table 4.8

Model Prediction Accuracy (R2) in the Validation Sample (N = 464) for Models Trained With Incrementally Restricted Training Sample Size

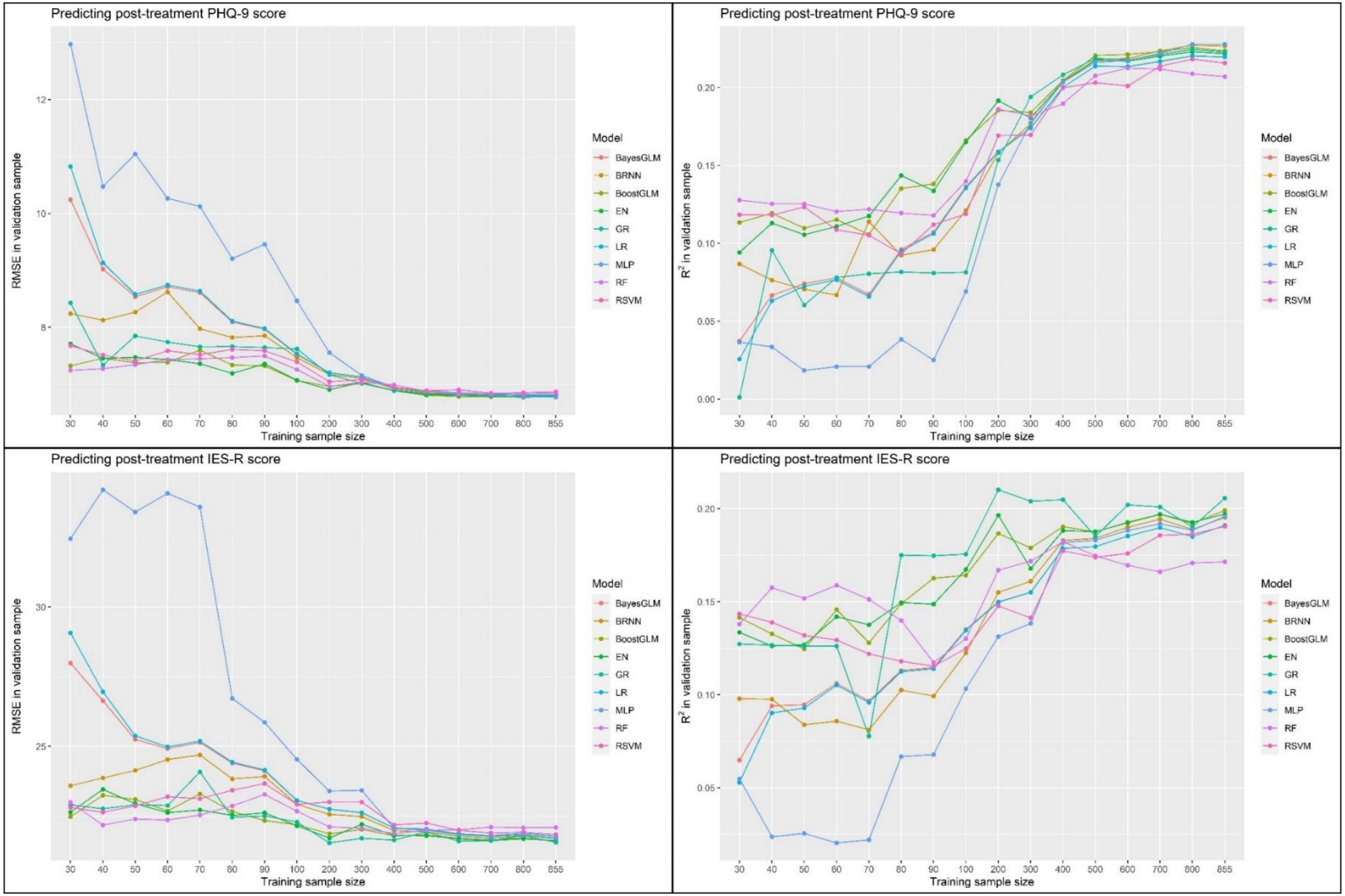
Outcome	Model	Training sample size															
		30	40	50	60	70	80	90	100	200	300	400	500	600	700	800	855
PHQ-9	LR	.026	.063	.072	.077	.066	.095	.106	.136	.159	.174	.200	.214	.213	.217	.220	.220
	BayesGLM	.037	.067	.074	.078	.067	.096	.107	.136	.159	.174	.200	.214	.213	.217	.220	.220
	BoostGLM	.113	.119	.110	.115	.106	**.135**	**.138**	**.166**	.185	.184	.205	**.221**	**.221**	.223	.226	.224
	BRNN	.087	.076	.070	.067	.114	.092	.096	.121	.158	.177	.204	.218	.219	**.223**	.227	.227
	EN	.094	.113	.105	.111	.117	.144	.134	.165	**.192**	.180	.204	.218	.217	.220	.223	.222
	GR	*.001*	.095	.060	.078	.080	.082	.081	.081	.153	**.194**	**.208**	.219	.218	.221	.225	.223
	MLP	.036	*.033*	*.018*	*.021*	*.021*	*.038*	*.025*	*.069*	*.138*	.177	.204	.216	.217	.222	**.228**	**.228**
	RF	**.128**	**.125**	**.125**	**.120**	**.122**	.119	.118	.140	.186	.182	*.190*	.208	.212	*.212*	*.209*	*.207*
	RSVM	.118	.118	.123	.109	.105	.093	.112	.119	.169	*.170*	.200	*.203*	*.201*	.214	.218	.216
IES-R	LR	*.053*	.090	.093	.105	.096	.112	.114	.135	.150	.155	.179	.180	.185	.190	.185	.191
	BayesGLM	.065	.094	.095	.106	.097	.113	.115	.135	.150	.155	.179	.180	.185	.190	.185	.191
	BoostGLM	.141	.133	.125	.146	.128	.149	.163	.164	.187	.179	.190	.188	.193	.197	.192	.199
	BRNN	.098	.098	.084	.086	.081	.103	.099	.122	.155	.161	.183	.184	.190	.194	.189	.195
	EN	.134	.126	.127	.142	.138	.150	.149	.167	.196	.168	.188	**.188**	.192	.197	**.193**	.197
	GR	.127	.127	.126	.126	.078	**.175**	**.175**	**.176**	**.210**	**.204**	**.205**	.185	**.202**	**.201**	.190	**.206**
	MLP	.055	*.024*	*.025*	*.020*	*.022*	*.067*	*.068*	*.103*	*.131*	*.138*	.182	.183	.188	.192	.188	.196
	RF	.138	**.158**	**.152**	**.159**	**.151**	.140	.117	.130	.167	.172	.182	.175	*.170*	*.166*	*.171*	*.171*
	RSVM	**.144**	.139	.132	.129	.122	.118	.115	.125	.148	.141	*.177*	*.174*	.176	.186	.186	.190

Note. The best performing model in each training sample is highlighted in **bold** and the worst is highlighted in *italic*.
Values are given to three decimal places to allow comparison of small differences in model performance.
IES-R = Impact of Event Scale – Revised; PHQ-9 = Patient Health Questionnaire 9; RMSE = Root Mean Squared Error.

Model: LR = Linear Regression; BayesGLM = Bayesian Generalised Linear Model; BoostGLM = Boosted Generalised Linear Model; BRNN = Bayesian Regularised Neural Network; EN = Elastic Net; GR = Genetic Regression; MLP = Multi-Layer Perceptron; RF = Random Forest; RSVM = Radial basis function Support Vector Machine.

Figure 4.4

Line Graphs Displaying Prediction Model Performance Metrics by Training Sample Size, for Models Predicting Post-treatment Score in the Validation Sample



Note. IES-R = Impact of Event Scale – Revised; PHQ-9 = Patient Health Questionnaire 9; RMSE = Root Mean Square Error.

Model: BayesGLM = Bayesian Generalised Linear Model; BoostGLM = Boosted Generalised Linear Model; BRNN = Bayesian Regularised Neural Network; EN = Elastic Net; GR = Genetic Regression; LR = Linear Regression; MLP = Multi-Layer Perceptron; RF = Random Forest; RSVM = Radial basis function Support Vector Machine.

4.4 Discussion

4.4.1 Summary of Main Findings

This study compared the performance of eight different ML methods and ordinary linear regression at predicting the outcome of Tf-CBT from pre-treatment data in routine clinical practice at NHS Talking Therapies services. For comparison with earlier studies, models were trained once with post-treatment PHQ-9 score as the outcome measure, and then again with post-treatment IES-R score as the outcome measure. There were only marginal (R^2 < .01, RMSE < .1) differences in performance between all models except for RF, which appeared to perform best in internal cross-validation but performed worst in external validation. RF exhibited considerably greater shrinkage than other models for both RMSE and R^2 and both outcome measures, suggesting the RF was overfit to the training sample. The variable importance metrics suggest that RF attributed greater importance to GAD-7 and WSAS than the linear models (LR, GR, EN, BoostGLM, BayesGLM), it is possible that RF modelled spurious non-linear relationships between these variables that did not generalise beyond the training sample. RF was the only non-linear model for which variable importance metrics were available; this is a limitation of the MLP, BRNN, and RSVM models. As with all ML methods, these methods prioritise accurate prediction in new data over explanation of relationships in current data. But, in the context of precision treatment selection, explainability may be important to both patients and clinicians. Calibration plots showed that the accuracy of all models was limited when predicting outcomes at the extreme ends of the scales, this is congruent with findings in Chapter 3 and seems to be a limitation of predicting treatment outcomes on a continuous scale.

4.4.1.1 Important Predictor Variables

There was some consensus between the linear models (LR, GR, BayesGLM, BoostGLM, EN) that patients with more severe PTSD and depression who were unemployed and of a younger age were less likely to respond to Tf-CBT. Taking anti-depressant medication was associated with higher post-treatment level of depression, but not with post-treatment level of PTSD. These might be factors that therapists could pay attention to and respectfully discuss with patients. Whether patients with these characteristics would be more or less likely to respond to another trauma focussed psychological therapy, such as EMDR, remains open to question. The finding that IMD was not among the most important predictors was unexpected given previous findings that IMD predicts PTSD treatment outcome in this dataset (Delgadillo & Richardson, 2024). However, it is possible that unemployment is a proxy indicator of social deprivation, and these two variables explain some of the same variance in treatment outcome (Spearman's correlation in training sample before imputation = -.31, see Appendix J). It is important to remember that these are multivariable models, and the coefficients represent the effect of each variable adjusted for all other variables in the model. Therefore, caution is advised when interpreting the effect of any individual variable. As these are linear models they do not account for non-linear relationships or interactions between variables, therefore some variables may be related to outcomes in ways that are not captured by these models. However, this doesn't seem to have negatively affected their prediction performance when compared to the non-linear models (BRNN, MLP, RF, and RSVM).

4.4.1.2 Training Sample Size

When comparing model performance across training sample sizes, again the differences were small. RMSE and R^2 for all models on both outcome measures began to converge and plateau once the training sample size exceeded $N = 400$. MLP performed exceptionally poorly with small training samples but performed marginally better than the other models on both metrics when predicting post-treatment PHQ-9 with larger training samples. This aligns with the estimation by Alwosheel et al. (2018) that 50 outcome events per parameter are required to train neural networks. It is possible that if the training sample size were increased further, MLP performance would continue to improve whereas the performance of the other models would plateau (Ng et al., 2020). However, MLP performance appears to begin to plateau at around the same point as the other models (Figure 4.3), and Giesemann et al. (2023) found that MLP performance peaked with a training sample size between 300 and 500, and increasing training sample size to 1000 and beyond did not improve ML model performance overall.

Conversely, although RF performed considerably worse than the other models with larger sample sizes, it was among the best performing models with smaller training samples and predicted both outcomes with greater accuracy than LR until training sample size reached $N = 100$, and less error than LR until $N = 400$. EN and RSVM also consistently performed better than LR with smaller training samples ($N < 100$), counter to simulation study findings that these methods require larger quantities of training data to train reliable models (Riley et al., 2021; van der Ploeg et al., 2014). The finding that model performance began to stabilise once the training sample size reached $N = 400$ is congruent with the findings of Giesemann et al. (2023), and adds empirical support to the minimum sample size calculation ($N = 337$, section 4.2.4) and the more conservative rule-of-thumb that 20 outcome events per candidate predictor variable parameter are required to train reliable clinical prediction models (see section 2.4.2).

4.4.2 Strengths and Limitations

The development of any clinical prediction model is naturally limited by the data available to train and test the model. In the time since the data used for this study was collected, NHS Talking Therapies have changed PTSD outcome measure from the IES-R to the PTSD Checklist for DSM-5 (PCL-5; Blevins et al., 2015). This potentially limits the extent to which these findings are applicable to current NHS Talking Therapies data and practice (NHS Digital, 2021). However, the similarity in the patterns of results for models trained to predict PHQ-9 and IES-R suggests that the methodological recommendations and important predictor variables reported here apply across measures of PTSD treatment outcome. The dataset used for this study lacked data on patients who received EMDR, precluding comparison of different ML methods capabilities to predict differential treatment response to Tf-CBT versus EMDR. As with the study reported in Chapter 3, only total scores were available on the clinical psychometric measures, and item level scores may be better predictors of outcome (Delamain et al., 2024).

A strength of this dataset compared to that used in Chapter 3 is that the IES-R was more consistently collected. However, the current dataset only included patients with a pre-treatment IES-R score, which may limit the representativeness of the sample. There was also a large proportion of missing values on post-treatment IES-R score, which were imputed with missForest. Stekhoven and Bühlmann (2012), and Waljee et al. (2013) demonstrated that missForest imputation error is lower than that of other imputation methods and is stable with up to 30% missing data, but larger proportions of missing data were not tested. Therefore, it is possible that missingness on the outcome variable introduced some bias into the IES-R prediction models. However, there were strong correlations between post-treatment IES-R score and post-treatment scores on PHQ-9, GAD-7 and WSAS, which were used for imputation; out-of-bag error estimates were similar to those reported in Chapter 3; and R^2

values for models predicting post-treatment IES-R were similar to those predicting post-treatment PHQ-9.

Following previous methodological recommendations, this study applied bootstrapping for internal cross-validation, and the .632 method was chosen as it is the most advanced method available in the caret package (Efron, 1983). It is possible that alternative bootstrapping methods (or using *k*-fold with multiple repetitions) could have improved model performance (Iba et al., 2021; Tantithamthavorn et al., 2017). But, in testing 9 prediction methods, 16 training sample sizes, and 2 outcome measures, this study compared multiple performance metrics for 288 models (not including models compared during grid search). Comparing different internal cross-validation methods in addition to this would have led to *combinatorial explosion* (i.e., rapid multiplicative increase in the number of models tested). Similarly, there is evidence that *random search* may be a more effective and efficient method of hyperparameter setting than grid search (Bergstra & Bengio, 2012; Bischl et al., 2023), and that other imputation methods may be more reliable than missForest under certain conditions (Hong & Lynn, 2020); but using grid search and missForest aids comparability with similar studies, including many of those reviewed in Chapter 2 and those that guided the selection of ML methods for this study (cited in section 4.2.8.1). In designing this study, a balance was sought between thorough examination of methodological questions and interpretability of the results, both within the study and in the context of the wider literature.

Following best practice recommendations (Steyerberg, 2019), geographic validation was applied to evaluate the predictive performance of the models. This is a more rigorous test of model generalisability than randomly partitioning the data, as the training and validation samples contained data from different NHS Talking Therapies services, with different patients and therapists, in different geographic locations. However, it is possible that differences in the distributions of variables between geographic regions could limit the extent to which a model trained on data from one region generalises to data from another region. For example, if the training sample is drawn from a region with high levels of social deprivation, but the validation sample is from a region with low levels of social deprivation, then the relationship between social deprivation and treatment outcome may be modelled on social deprivation values that are out of range for the validation sample, and model accuracy in this sample will be impaired. In the current study, the training and validation samples each contained data from multiple NHS Trusts in different geographic locations, therefore the likelihood of this limitation was reduced. Additionally, due to prior evidence that pre-treatment PTSD and social deprivation are related to PTSD treatment outcome, Trusts were drawn at random until a validation sample was selected that was not significantly different to the training sample on pre-treatment IES-R score or IMD.

Nevertheless, the systematic review findings presented in Chapter 2 suggest that this study was the first to use ML methods to predict psychological therapy outcomes for PTSD [1] that performed a sample size calculation; [2] that had an adequate sample size; [3] that tested the effect of training sample size; [4] that tested Bayesian methods; [5] that tested deep learning methods; [6] that tested geographic validation; and [7] that compared the performance of different ML methods against each other and a non-ML comparator in a validation sample. Hence, best practice in ML were adhered to.

4.4.3 Recommendations for Future Research

Where models perform equally well in terms of prediction accuracy and error, other factors may guide predictor model selection such as the interpretability and explainability of the model. With training samples exceeding the minimum size recommended by the sample size calculation, the simpler linear models (LR, BayesGLM, BoostGLM, EN, GR) performed as well as the more complex, non-linear models (BRNN, MLP, RF, RSVM) and were more interpretable. Additionally, GR, EN, and BoostGLM perform predictor selection, further reducing model complexity and increasing interpretability, and these models performed marginally better than linear regression with the full training sample. If explainability is not a priority, and there is a large training sample available, the deep learning methods (MLP, BRNN) may offer slightly better prediction accuracy. Future qualitative research could provide insight into the importance of explainability to clinicians and patients in NHS Talking Therapies settings.

Where the training sample size is limited, MLP is not recommended (although other neural networks such as BRNN may perform better). However, counter to some recommendations (e.g., Moons et al., 2019), ML methods such as EN and BoostGLM appear to offer an even greater advantage over LR when the training sample size is small. This may be because their predictor selection and regularisation procedures reduce overfitting, as intended. LR, on the other hand, seems to require a much larger training sample than typically applied ($N = 400$, or 20 events per variable) to compete with these models. When the training sample is large, RF is not recommended, as it appears to overfit the training sample and produce the least accurate out-of-sample predictions. RF may offer some advantages when the training sample is small, but as with all of these methods it is important to test performance in an external validation sample as it can be overestimated in internal cross-validation. Sample size calculation is always recommended (Riley et al., 2020).

Future studies could investigate whether predicting a binary outcome measure, such as reliable change in symptoms, improves model calibration. This may also lead to more clinically useful prediction models, as rather than producing a predicted final session score on a symptom measure (with limited accuracy for outliers), the model would produce a predicted probability that an individual patient would respond to a particular treatment (likely to be more accurate for outliers at the extreme ends of the severity scale, but limited by some classification errors for cases in the boundary of the binary classifier's cut-off). Clinicians could use this information when deciding whether to pursue a particular treatment approach with a patient. A larger dataset would be required to test this as sample size calculations are based on the number of outcome events (i.e., the number of patients in the smallest outcome category), and only around a third of patients accessing Tf-CBT at NHS Talking Therapies services have a reliable and clinically significant improvement in symptoms (Robinson et al., 2020).

Future studies could compare the effect of different internal cross-validation methods (e.g., bootstrap .632 vs. 10-fold with 10 repetitions), different hyperparameter optimisation methods (e.g., grid search vs. random search), and/or different imputation methods (e.g., missForest vs. nearest neighbour imputation) on ML clinical prediction model performance metrics. Studies could also test whether ensemble methods, such as applying one ML method for predictor selection and another for prediction, and/or combining models for prediction (e.g., Bennemann et al., 2022), improve model performance in this context. All the above comments also apply to presenting problems beyond PTSD.

4.4.4 Conclusions

The findings of this study indicate that [1] some ML methods do predict the outcome of Tf-CBT with greater accuracy and less error than LR, [2] some ML methods are better than others at this task, and [3] differences in model performance are moderated by sample size. EN and BoostGLM performed consistently well across sample sizes, were considerably better than LR with smaller sample sizes, and marginally better with large sample sizes. RF and RSVM performed well with smaller sample sizes but were no better than LR with large sample sizes; whereas MLP performed notably poorly with small sample sizes but was among the best models when trained with the largest sample sizes. GR was less consistent with smaller sample sizes but (along with EN) performed best at predicting post-treatment PTSD with larger training sample sizes. A sample size calculation (Riley et al., 2020) is always recommended whether using ML methods or LR, and as a rule-of-thumb at least 20 outcome events per candidate predictor parameter are required. When predicting Tf-CBT outcomes in NHS Talking Therapies data, BoostGLM, EN, or GR are recommended to maximise prediction accuracy and explainability. External validation is essential as internal cross-validation procedures may not prevent overfitting.

CHAPTER 5
General Discussion

Approximately 4% of adults experience post-traumatic stress disorder (PTSD; Fear et al., 2016; Koenen et al., 2017), a condition that entails considerable suffering and functional impairment following exposure to trauma (Yehuda et al., 2015). Despite the availability of multiple effective psychological therapies (Jericho et al., 2021), PTSD has lower treatment response rates than other mental health problems (Robinson et al., 2020). Precision treatment selection using machine learning (ML) has the potential to improve both clinical effectiveness and service efficiency (Chekroud et al., 2021). However, relatively little precision treatment research has focussed on the treatment of PTSD (Aafjes-van Doorn et al., 2021; Vieira et al., 2022). Therefore, the aim of this thesis was to advance precision treatment selection for PTSD, with a particular focus on NHS Talking Therapies services. NHS Talking Therapies are well suited to advancing data driven approaches to treatment, as they are free at the point of access, treat over 66,000 patients with PTSD per year (NHS Digital, 2024), and collect a standardised dataset across services (NHS Digital, 2021).

5.1 Summary of Findings

Chapter 2 presented the first systematic review of studies that applied ML methods to predict the outcome of clinical practice guideline (CPG) recommended psychological therapies for PTSD. In addition to assessing risk of bias (Moons et al., 2019), the methodological rigour of each study was assessed against the best practice ML pipeline (Delgadillo & Atzil-Slonim, 2022). Seventeen studies met the inclusion criteria for the review, and sixteen were published in the preceding four years. All but one study reported identifying significant predictors of outcome, but all were at high risk of bias, none performed a sample size calculation, and only one study tested external validation. Four studies developed models for precision treatment selection, but none were externally validated.

One study (Deisenhofer et al., 2018) developed a personalised advantage index (PAI) to predict response to trauma-focussed cognitive behavioural therapy (Tf-CBT) and eye movement desensitisation and reprocessing (EMDR) in NHS Talking Therapies services. In internal cross-validation, there was a significantly higher rate of reliable improvement among patients who had received their model indicated optimal treatment compared to patients who had received their suboptimal treatment. This model was not externally cross validated, and Chapter 3 tested the external validity of this model by applying it in an independent sample of NHS Talking Therapies clinical case records. This study found that the models did not make treatment specific predictions and the PAI failed to identify the optimal treatment for individual patients. This suggested that the models may have been overfit to the training data, possibly due to the training sample being too small, the internal cross-validation procedure being insufficient, or the choice of ML method (genetic regression) being inappropriate for the task.

Chapter 4 explored this by comparing different ML methods at predicting the outcome of Tf-CBT, applying bootstrapping for rigorous internal cross-validation, performing a sample size calculation (estimated minimum required $N = 337$), and iteratively restricting the training sample size from $N = 855$ to $N = 30$. Performance was evaluated in a geographic validation sample (i.e., data from different NHS Trusts). This study found that once the training sample size exceeded the required size of $N = 337$, the choice of ML method made only a marginal difference to out-of-sample prediction performance. Boosted generalised linear model (BoostGLM) and elastic net (EN) were among the best performing models across sample sizes, measures, and metrics. These models are more transparent and interpretable than the deep learning and non-linear support vector machine models, and they produce parsimonious models by excluding unimportant predictors. With training samples greater than the recommended minimum *N* = 337 (approximately 20 events per candidate predictor parameter), BoostGLM and EN were marginally better than linear regression, and offered a greater advantage with smaller sample sizes. With the largest training sample (*N* = 855), Random Forest (RF) overfit the training sample to a much greater degree than all of the other models, indicated by the considerable shrinkage from internal cross-validation to external validation performance. Interestingly, genetic regression (GR) was the best model at predicting post-treatment PTSD (measured with the IES-R) in the validation sample, with the lowest prediction error (RMSE = 21.56), highest accuracy (R^2 = .206), and second lowest shrinkage after multi-layer perceptron (MLP).

5.2 Interpretation of Results

Comparing the out-of-sample performance of the GR model predicting post-Tf-CBT PHQ-9 score in Chapter 3 (RMSE = 6.92, R^2 = .28), to that of the GR model predicting post-Tf-CBT PHQ-9 score in Chapter 4 (RMSE = 6.80, R^2 = .22), reveals better prediction accuracy for the model tested in Chapter 3 (and only marginally worse error). Hence, although the small validation samples in Chapter 3 reduce confidence in these estimates, it is unclear whether the choice of ML method or internal cross-validation method contributed to the failure of the PAI to predict the optimal treatment for individual patients in the validation sample in Chapter 3.

However, the RMSE and R^2 values do not measure the extent to which the models make treatment specific predictions, which is the assumption underlying the PAI. Kessler et al. (2017) suggested training a separate prognostic model for each treatment group to address the problem that most psychological therapy samples are underpowered to test interaction effects. However, this assumes that training a separate model for each treatment group will produce predictions that are specific to the respective treatment, but the most important predictor variables are likely to predict outcomes irrespective of psychological therapy type. Although the pattern of R^2 and RMSE values in the development sample (Table 3.3) indicated that the PAI made predictions that were specific to the respective treatment *group* (i.e., the individual patients), the pattern of R^2 and RMSE values in the validation sample indicated that the predictions were not specific to the respective treatment *type* (i.e., the therapeutic mechanisms of action). The alternative approach would be to acquire a large enough sample to train a single prescriptive model that includes treatment type as a variable and identifies patient characteristics that interact with treatment type, i.e., aptitude-by-treatment interactions (ATI; Cronbach & Snow, 1977). This was the approach taken in the early PAI studies (DeRubeis et al., 2014; Huibers et al., 2015), but these studies did not use ML methods and did not test external validity.

Delgadillo and Gonzalez Salas Duhne (2020) compared the predictive accuracy of treatment specific prognostic models to that of an ATI model predicting outcomes in a sample of patients that received CBT or counselling for depression, and found that the separate prognostic models demonstrated less shrinkage in a randomly partitioned validation sample. This is coherent with the findings presented in Chapter 4, as the prognostic models were EN and the ATI model was RF. However, unlike the findings of Chapter 3, Delgadillo and Gonzalez Salas Duhne (2020) found that their PAI generalised to their validation sample, in that there was a significant difference in the rate of reliable improvement between patients who received their model indicated optimal and suboptimal treatments.

There are three key methodological differences that may explain why the PAI developed by Delgadillo and Gonzalez Salas Duhne (2020) generalised to the validation sample, but the PAI developed by Deisenhofer et al. (2018) did not: [1] It could be that EN is better able to model treatment specific predictors than GR; [2] a larger training sample may be required to develop treatment specific prediction models, Delgadillo and Gonzalez Salas

Duhne (2020) training sample included *n* = 929 patients who had received CBT and *n* = 156 who had received counselling, whereas Deisenhofer et al. (2018) included *n* = 150 patients who had received Tf-CBT and *n* = 75 who had received EMDR; [3] Delgadillo and Gonzalez Salas Duhne (2020) trained models to predict reliable change in symptoms as a binary outcome, rather than final session symptom score, and it is possible that treatment specific predictors of reliable change are easier to identify than treatment specific predictors of post-treatment symptom severity.

However, another possible reason that the PAI developed by Delgadillo and Gonzalez Salas Duhne (2020) successfully predicted the optimal treatment for individual patients in a validation sample, whereas the PAI developed by Deisenhofer et al. (2018) did not, is that CBT and counselling for depression have very different mechanisms of action, whereas Tf-CBT and EMDR do not. If so, it may make more of a difference to patients with depression whether they receive CBT or counselling, whereas it does not make as much of a difference (if any at all) to patients with PTSD whether they receive Tf-CBT or EMDR. Alternatively, there may just be fewer patients with PTSD who respond differentially to Tf-CBT versus EMDR. Both these possibilities would mean that a larger sample would be required to identify the combination of patient characteristics that interact with response to TF-CBT versus EMDR. Recent PAI studies with external validation found that receiving the model indicated optimal treatment only improved outcomes for a minority of patients with the most robust PAI recommendations (Moggia et al., 2023; B. Schwartz et al., 2021). Additionally, as all NHS Talking Therapies high intensity therapists are trained in Tf-CBT, EMDR is most often delivered as part of an integrated treatment with Tf-CBT. As such, it may be more clinically useful to identify patients who are most likely to benefit from the addition of EMDR to Tf-CBT, for example by developing a PAI for Tf-CBT versus Tf-CBT+EMDR. Regardless, the findings of Chapter 3 are in line with recent studies testing the external validity of PAI, which suggest that earlier predictions about the potential for precision treatment selection using ML methods may have been somewhat over-optimistic (Moggia et al., 2023; Van Bronswijk et al., 2021). Further, PAI for psychological therapies are yet to be prospectively tested; therefore, it is unknown whether applying an externally validated PAI to prospectively assign patients to treatment will make any difference to PTSD treatment outcomes in practice.

Nevertheless, the findings in Chapter 4 indicate that up to 20% ($R^2 = .2$) of the variance in Tf-CBT outcome (measured with the IES-R) can be predicted at the point of assessment in routine care using data that NHS Talking Therapies services already collect. Compared to similar studies, this is relatively promising. Only one study in the systematic review in Chapter 2 tested the external validity of a ML model, they used EN to predict PTSD outcomes from routinely collected data from specialised inpatient clinics in Germany and found that the model explained 16% ($R^2 = .16$) of the variance in outcomes in the validation sample. Treatment selection aside, this prognostic information may be useful to patients, clinicians and service providers when planning treatment. The calibration plots in Figures 4.1 and 4.2 indicate that models were poor at predicting outcomes at the extreme ends of the distribution, and the average prediction error for the best models (see Table 4.8) was greater than the respective indices of reliable improvement for the PHQ-9 (≥ 6; Richards & Borglin, 2011) and IES-R (≥ 9; NHS England, 2014). However, clinical prediction models do not need to be perfect to be clinically useful, they only need to out-perform clinical intuition. R. Schwartz et al. (2021) found that most variables that clinicians perceived to be predictors of CBT outcome were not associated with outcome. Indeed, the one variable endorsed by clinicians that was significantly associated with outcome, was associated with change in the opposite direction to that predicted by clinicians. This is entirely congruent with the history of clinical versus statistical judgement research (Ægisdóttir et al., 2006; Meehl, 1954).

5.3 Clinical and Service Implications

There was some consensus among the models trained in Chapter 4 that patients with more severe PTSD and depression, who were unemployed, and of younger age, had more severe PTSD post-treatment. Although age appeared to be less important than the other three predictors, these findings indicate that patients with this combination of factors may be less likely to respond to Tf-CBT in NHS Talking Therapies. This information may be useful to clinicians when planning treatment for a patient with PTSD, as patients meeting this description may require more sessions of Tf-CBT, or may benefit more from another treatment, such as EMDR. Unfortunately, the lack of EMDR data in the Chapter 4 dataset precluded investigation of this.

There are several caveats to using this information to guide clinical decision-making. Firstly, it is important to remember that these are multivariable models, which suggest that the cumulative effect of these variables may be related to poorer outcomes, but this does not necessarily indicate that any of these variables in isolation significantly predict poorer outcomes. Secondly, these variables may be proxy indicators of other variables that were not measured in the dataset. For example, unemployment could be associated with various forms of poverty and deprivation, and younger age could be associated with more recent onset of PTSD or experiencing trauma earlier in life. If this were the case, it may not be unemployment and age that lead to poorer Tf-CBT outcomes, but the unmeasured variables with which they correlate. Thirdly, external validation of prediction models does not directly assess the external validity of the associations between individual predictors and outcome, replication with methods designed for explanatory modelling (e.g., structural equation modelling) would provide stronger evidence for these relationships. The goal of ML methods is to make predictions, rather than establish causal relationships between predictors and outcome. The systematic review in Chapter 2 found that the important predictors identified by studies were heterogenous, but pre-treatment PTSD and depression were among the most frequently selected predictors, and age was selected by a number of studies. This provides some support for the important predictors identified in Chapter 4.

However, an important finding of the systematic review in Chapter 2 was that all studies were at high risk of bias, none adhered to every step of the ML pipeline, and only one study provided level 3 evidence (external validity). Furthermore, Chapters 3 and 4 highlight the importance of external validation. As such, caution is advised when applying these findings in clinical practice. It is important to critically evaluate ML modelling studies paying particular attention to sample size calculation, number of outcome events per predictor parameter, and the level of evidence (apparent validation, internal cross-validation, external validation).

Effectively matching patients to the therapy most likely to enable recovery would mean that treatment outcomes improve for patients and wait-times are reduced for services, as fewer patients return for treatment (Lorimer et al., 2024). Services may also need to ensure that regular checking of therapist competency is assessed, and this could be part of the minimum dataset for NHS Talking Therapies services.

5.4 General Strengths and Limitations of the Thesis

5.4.1 Strengths

The focus on PTSD has advanced understanding of how to better personalise treatment for this condition. Rigorous methods were used, including a systematic review that followed appropriate best practice guidelines (Moons et al., 2014, 2019; Page et al., 2021), external validation in a field that is currently lacking in this level of evidence, and cutting edge ML prediction methods such as deep learning and Bayesian methods that have not yet been applied in this field. Other cutting-edge quantitative methods included a non-parametric ML approach to multiple imputation, and propensity score matching to address confounding by indication in routine clinical practice data. These methods were applied and reported following current best practice guidelines (Collins et al., 2024). The empirical research benefited from the use of NHS Talking Therapies data, which provided large, naturalistic datasets collected in routine practice that have high ecological validity. Chapter 4 included the largest sample of PTSD treatment outcome data utilised in ML prediction modelling to date, and models were developed and tested in the context of their intended use. This means that the findings are directly applicable to NHS Talking Therapies services, and generalisability was demonstrated through geographic validation. Novel contributions to the field included the first systematic review of the use of ML methods to predict the outcome of psychological therapies for PTSD, the first external validation of a PAI for PTSD, and the first robust exploration of ML methods for predicting the outcome of psychological therapy for PTSD in NHS Talking Therapies.

5.4.2 Limitations

The empirical research relied on secondary analysis of routinely collected data. As such, there was no control over the variables in the dataset, and variables that were frequently identified as predictors of PTSD outcome in the systematic review, such as trauma related variables, could not be added to the dataset. At present, the NHS Talking Therapies dataset only aggregates total scores on the clinical measures, precluding investigation of specific symptoms as predictors or moderators of treatment outcome. A number of studies in the systematic review in Chapter 2 found that item level or subscale scores were important predictors (e.g., Held et al., 2022; Herzog et al., 2021), and Delamain et al. (2024) found that item level scores on the PHQ-9 and GAD-7 were more important predictors than total scores when predicting anxiety outcomes in NHS Talking Therapies data. There is evidence that subtypes of depression, identifiable through analysis of item-level scores on the PHQ-9, respond differentially to CBT (Catarino et al., 2022; Simmonds-Buckley et al., 2021). PTSD is similarly clinically heterogeneous (Galatzer-Levy & Bryant, 2013) and it is possible that subtypes of PTSD (e.g., Campbell et al., 2020) respond differentially to Tf-CBT and/or EMDR.

The dataset used in Chapter 3 had so little EMDR data that the sample was much smaller than recommended. There was also substantial missing data on some variables in this dataset, to the extent that the IES-R could not be used as an outcome measure. The dataset used in Chapter 4 lacked data on patients treated with EMDR, precluding investigation of moderators and development of a PAI. Although this limits the extent to which this study advances precision treatment selection for PTSD, the comparison of different ML methods and training sample sizes provides important methodological foundations upon which to do so. The training sample in Chapter 4 was much larger than any in the systematic review in Chapter 2, but some ML methods such as MLP may require even larger datasets for peak performance. As this was routine practice data, patients were not randomised to treatments, introducing potential confounding by indication. Propensity score matching was used to balance baseline differences in observed covariates, but this cannot balance differences in unobserved covariates in the same way that randomisation does. Treatment delivery is not as rigorously controlled in routine practice as in clinical trials, there are no fidelity or competency checks, and clinicians are relied upon for data collection. However, all NHS Talking Therapies high intensity therapists receive the appropriate training and practice under regular clinical supervision.

All the empirical research in this thesis was quantitative, and qualitative research is required to understand patients and therapists' perspectives on the acceptability and usefulness of precision treatment selection. Empirical studies focussed on the English NHS Talking Therapies context and may not generalise to different treatment contexts or different psychometric measures. Chapter 4 findings may not generalise to trauma-focussed therapies other than Tf-CBT, and analyses using the IES-R may not generalise to the current PTSD measure used by NHS Talking Therapies services, the PTSD Checklist for DSM-5 (PCL-5). All clinical measures were self-report and may be vulnerable to biases such as social desirability bias and response bias. A significant proportion of variance in psychological therapy outcomes is due to differences between therapists and differences between services (Firth et al., 2019), and this was not controlled for in the analyses. None of the evidence presented in the empirical studies included follow-up and therefore issues of clinical durability versus relapse were unexamined.

5.5 Recommendations for Future Research Directions

To provide a robust proof of concept for a PAI for Tf-CBT versus EMDR for PTSD in NHS Talking Therapies would require a large sample of clinical case records with sufficient data from patients who had received EMDR, and patient outcomes measured with the PCL-5. The exact sample size could be determined by a sample size calculation and would depend on a number of factors described below. Ideally, the data would be split geographically by NHS Trust into three samples: training, validation, and test sample. A range of different ML methods could be used to develop a PAI in the training sample and compared in the validation sample. Models predicting reliable improvement as a binary outcome could be compared to models predicting final session score as a continuous outcome, to investigate whether binarising outcome improves prediction error and calibration. Both methods of developing PAI could be compared: separate prognostic models for Tf-CBT and EMDR, and single prescriptive (ATI) models that include treatment type as a moderator variable. The best performing PAI model can then be retrained on the combined training and validation sample and subjected to a final test of external validity in the test sample. The sample size calculation in Chapter 4 found that a minimum sample size *N* = 337 was required for the training sample to predict continuous outcomes. As only around a third of patients accessing Tf-CBT at NHS Talking Therapies have a reliable improvement in symptoms (Robinson et al., 2020), this would correspond to training sample of around *N* = 1,000 patients to predict reliable improvement (and *N* = 750 for the validation and test samples respectively). This suggests that a sample size of approximately *N* = 2,500 per treatment (*N* = 5,000 in total) would be sufficient for robust development, validation and testing of a PAI for Tf-CBT versus EMDR in NHS Talking Therapies. By the usual standards of psychological therapy research, this would be an enormous sample size, but NHS Talking Therapies services across England treat over 66,000 patients with PTSD each year (NHS Digital, 2024), therefore acquiring a sample of clinical case records this size would not be impossible.

With the caveat that adding parameters increases the required sample size, there are several variables that may improve prediction accuracy or otherwise strengthen the PAI if they were added to the model. These include item-level scores on the relevant clinical measures (PCL-5, PHQ-9 and GAD-7), particularly PTSD symptoms or symptom clusters, and trauma-related variables such as type of trauma, time since trauma, and post-traumatic cognitions (as discussed in section 2.3.4.3). Further, evidence that the therapies had been delivered as intended may improve the likelihood of the PAI being successful, such as checks that treatment protocols were adhered to and Tf-CBT sessions included work on traumatic memories.

Following successful external validation, the PAI can be prospectively tested by applying it in NHS Talking Therapies services to assign patients to Tf-CBT or EMDR at the point of access. In an initial pilot study, patients’ outcomes could be compared to propensity score matched historical controls from NHS Talking Therapies clinical case records, and acceptability and feasibility explored. A subsequent RCT could randomise patients to PAI guided treatment selection versus treatment selection as usual to test the efficacy of the PAI. As a further test of the clinical usefulness of the model, therapists could be asked to predict patients' outcomes, and the model's prediction error compared to the therapists' prediction error.

An important next step is to work with key stakeholders to co-produce digital technologies to implement treatment selection algorithms in a way that is acceptable to patients, therapists, and services. Patients' and therapists' attitudes towards PAI guided treatment selection could be investigated by integrating a qualitative study within a pilot study. If patients and therapists do not view the PAI as useful or trustworthy, then therapists may be unlikely to use it, patients may be more likely to dropout from PAI allocated treatment, and the PAI will not be effective in routine practice. Therefore, it is essential to involve patients, therapists, and experts by experience in the development and implementation of precision treatment tools (Deisenhofer et al., 2024). Additionally, it is likely that some patients who receive their model-indicated optimal treatment still will not benefit from treatment, and qualitative research could explore the reasons for this.

This thesis provides several methodological recommendations for clinical prediction research more broadly. Whether using ML or linear regression, future studies looking to predict psychological therapy outcomes should always perform a sample size calculation to estimate the minimum required training sample size, and test model performance in a validation sample. There are R packages available that facilitate sophisticated sample size calculation for this purpose (e.g., pmsampsize). ML methods such as EN and BoostGLM offer an advantage offer LR, especially with smaller training sample sizes, and these are linear models that allow some degree of interpretability.

5.6 Conclusion

This thesis aimed to advance precision treatment selection for PTSD to improve the effectiveness of routinely delivered trauma-focussed psychological therapies. The findings identified significant limitations in previous precision PTSD treatment research using ML methods but identified a number of advantages of using ML methods to predict outcomes when following best practice guidelines. Precision treatment selection for PTSD may yet be possible, best practice guidelines are available to guide this future research and open access software is available to facilitate it. But, thus far, ML methods have not been applied reliably in adequate clinical datasets. Patients with PTSD will likely be better served by precision treatment in the long term when studies are designed and conducted with high compliance to best practice guidelines in the short and median term.

REFERENCES

Aafjes-van Doorn, K., Kamsteeg, C., Bate, J., & Aafjes, M. (2021). A scoping review of machine learning in psychotherapy research. *Psychotherapy Research*, *31*(1), 92–116. https://doi.org/10.1080/10503307.2020.1808729

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., Cohen, G., & 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. https://doi.org/10.1177/0011000005285875

Alwosheel, A., van Cranenburgh, S., & Chorus, C. G. (2018). Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *Journal of Choice Modelling*, *28*, 167–182. https://doi.org/10.1016/j.jocm.2018.07.002

American Psychiatric Association. (1952). *Diagnostic and statistical manual of mental disorders* (1st ed.). Author.

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Author.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Author.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Author.

American Psychiatric Association. (2022). *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision*. Author.

American Psychological Association. (2017). *Clinical practice guideline for the treatment of posttraumatic stress disorder (PTSD) in adults*. Author. https://www.apa.org/ptsd-guideline

Bækkelund, H., Endsjø, M., Peters, N., Babaii, A., & Egeland, K. (2022). Implementation of evidence-based treatment for PTSD in Norway: Clinical outcomes and impact of probable complex PTSD. *European Journal of Psychotraumatology*, *13*(2), 2116827. https://doi.org/10.1080/20008066.2022.2116827

Balki, I., Amirabadi, A., Levman, J., Martel, A. L., Emersic, Z., Meden, B., Garcia-Pedrero, A., Ramirez, S. C., Kong, D., Moody, A. R., & Tyrrell, P. N. (2019). Sample-Size Determination Methodologies for Machine Learning in Medical Imaging Research: A Systematic Review. *Canadian Association of Radiologists Journal*, *70*(4), 344–353. https://doi.org/10.1016/j.carj.2019.06.002

Baranyi, G., Cassidy, M., Fazel, S., Priebe, S., & Mundt, A. P. (2018). Prevalence of Posttraumatic Stress Disorder in Prisoners. *Epidemiologic Reviews*, *40*(1), 134–145. https://doi.org/10.1093/epirev/mxx015

Barawi, K. S., Lewis, C., Simon, N., & Bisson, J. I. (2020). A systematic review of factors associated with outcome of psychological treatments for post-traumatic stress disorder. *European Journal of Psychotraumatology*, *11*(1), 1774240. https://doi.org/10.1080/20008198.2020.1774240

Barber, J. P., & Muenz, L. R. (1996). The role of avoidance and obsessiveness in matching patients to cognitive and interpersonal psychotherapy: Empirical findings from the treatment for depression collaborative research program. *Journal of Consulting and Clinical Psychology*, *64*(5), 951.

Barkham, M. (2023). Smaller effects matter in the psychological therapies: 25 years on from Wampold et al. (1997). *Psychotherapy Research*, *33*(4), 530–532. https://doi.org/10.1080/10503307.2022.2141589

Beck, A. T., & Emery, G. (1985). *Anxiety Disorders and Phobias: A Cognitive Perspective*. Basic Books.

Ben–Ezra, M. (2004). Trauma in antiquity: 4000 year old post-traumatic reactions?. *Stress and Health: Journal of the International Society for the Investigation of Stress*, *20*(3), 121-125. https://doi.org/10.1002/smi.1003

Bennemann, B., Schwartz, B., Giesemann, J., & Lutz, W. (2022). Predicting patients who will drop out of out-patient psychotherapy using machine learning algorithms. *The British Journal of Psychiatry*, *220*(4), 192–201. https://doi.org/10.1192/bjp.2022.17

Berger, W., Coutinho, E. S. F., Figueira, I., Marques-Portella, C., Luz, M. P., Neylan, T. C., Marmar, C. R., & Mendlowicz, M. V. (2012). Rescuers at risk: A systematic review and meta-regression analysis of the worldwide current prevalence and correlates of PTSD in rescue workers. *Social Psychiatry and Psychiatric Epidemiology*, *47*(6), 1001–1011. https://doi.org/10.1007/s00127-011-0408-2

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, *13*, 281–305.

Bi, Q., Goodman, K. E., Kaminsky, J., & Lessler, J. (2019). What is Machine Learning? A Primer for the Epidemiologist. *American Journal of Epidemiology*, *188*(12), 2222–2239. https://doi.org/10.1093/aje/kwz189

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A.-L., Deng, D., & Lindauer, M. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *WIREs Data Mining and Knowledge Discovery*, *13*(2), e1484. https://doi.org/10.1002/widm.1484

Blackmore, R., Boyle, J. A., Fazel, M., Ranasinha, S., Gray, K. M., Fitzgerald, G., Misso, M., & Gibson-Helm, M. (2020). The prevalence of mental illness in refugees and asylum seekers: A systematic review and meta-analysis. *PLoS Medicine*, *17*(9). https://doi.org/10.1371/journal.pmed.1003337

Blevins, C. A., Weathers, F. W., Davis, M. T., Witte, T. K., & Domino, J. L. (2015). The Posttraumatic Stress Disorder Checklist for DSM-5 (PCL-5): Development and Initial

Psychometric Evaluation. *Journal of Traumatic Stress*, *28*(6), 489–498. https://doi.org/10.1002/jts.22059

Bone, C., Simmonds-Buckley, M., Thwaites, R., Sandford, D., Merzhvynska, M., Rubel, J., Deisenhofer, A.-K., Lutz, W., & Delgadillo, J. (2021). Dynamic prediction of psychological treatment outcomes: Development and validation of a prediction model using routinely collected symptom data. *The Lancet Digital Health*, *3*(4), e231–e240. https://doi.org/10.1016/S2589-7500(21)00018-2

Breiman, L. (2001a). Random forests. *Machine Learning*, *45*, 5–32. https://link.springer.com/article/10.1023/a:1010933404324

Breiman, L. (2001b). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Breuer, J., & Freud, S. (1893). On The Psychical Mechanism of Hysterical Phenomena. In J. Strachey (Ed.), *Standard edition of the complete psychological works of Sigmund Freud* (Vol. 2). Hogarth Press. https://www.themoralinjuryinstitute.com/wp-content/uploads/2020/01/On-the-Psychical-Mechanism-of-Hysteria-Preliminary-Communication.pdf

Burgess, A. W., & Holmstrom, L. L. (1974). Rape Trauma Syndrome. *American Journal of Psychiatry*, *131*(9), 981–986. https://doi.org/10.1176/ajp.131.9.981

Campbell, Sarah. B., Trachik, B., Goldberg, S., & Simpson, Tracy. L. (2020). Identifying PTSD symptom typologies: A latent class analysis. *Psychiatry Research*, *285*, 112779. https://doi.org/10.1016/j.psychres.2020.112779

Catarino, A., Fawcett, J. M., Ewbank, M. P., Bateup, S., Cummins, R., Tablan, V., & Blackwell, A. D. (2022). Refining our understanding of depressive states and state transitions in response to cognitive behavioural therapy using latent Markov modelling. *Psychological Medicine*, *52*(2), 332–341. https://doi.org/10.1017/S0033291720002032

Cearns, M., Hahn, T., & Baune, B. T. (2019). Recommendations and future directions for supervised machine learning in psychiatry. *Translational Psychiatry*, *9*(1), 1–12. https://doi.org/10.1038/s41398-019-0607-2

Chekroud, A. M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., Cohen, Z., Belgrave, D., DeRubeis, R., Iniesta, R., Dwyer, D., & Choi, K. (2021). The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry*, *20*(2), 154–170. https://doi.org/10.1002/wps.20882

Chen, Z. S., Kulkarni, P. (Param), Galatzer-Levy, I. R., Bigio, B., Nasca, C., & Zhang, Y. (2022). Modern views of machine learning for precision psychiatry. *Patterns*, *3*(11), 100602. https://doi.org/10.1016/j.patter.2022.100602

Clark, D. M., Layard, R., Smithies, R., Richards, D. A., Suckling, R., & Wright, B. (2009). Improving access to psychological therapy: Initial evaluation of two UK demonstration sites. *Behaviour Research and Therapy*, *47*(11), 910–920.

Clark, D. M., & Whittington, A. (2023, January 16). *What's in a name? NHS Talking Therapies, for anxiety and depression – the new name for IAPT services*. NHS England. https://www.england.nhs.uk/blog/whats-in-a-name-nhs-talking-therapies-for-anxiety-and-depression-the-new-name-for-iapt-services/

Cloitre, M., Courtois, C., Ford, J., Green, B., Alexander, P., Briere, J., & Van der Hart, O. (2012). *The ISTSS expert consensus treatment guidelines for complex PTSD in adults*. https://psychotraumanet.org/sites/default/files/documents/Cloitre-ISTSS%20Expert%20Consensus%20Guidelines%20for%20Complex%20PTSD.pdf

Cohen, Z. D. (2018). *Treatment Selection: Understanding What Works for Whom in Mental Health* [Ph.D., University of Pennsylvania]. https://www.proquest.com/docview/2117235776/abstract/3CA5132CE7244197PQ/1

Cohen, Z. D., Delgadillo, J., & DeRubeis, R. J. (2021). Personalized treatment approaches. In M. Barkham, W. Lutz, & L. G. Castonguay (Eds.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (7th ed., pp. 673–704). Wiley.

Cohen, Z. D., & DeRubeis, R. J. (2018). Treatment Selection in Depression. *Annual Review of Clinical Psychology*, *14*(1), 209–236. https://doi.org/10.1146/annurev-clinpsy-050817-084746

Cohen, Z. D., Kim, T. T., Van, H. L., Dekker, J. J. M., & Driessen, E. (2020). A demonstration of a multi-method variable selection approach for treatment selection: Recommending cognitive–behavioral versus psychodynamic therapy for mild to moderate adult depression. *Psychotherapy Research*, *30*(2), 137–150. https://doi.org/10.1080/10503307.2018.1563312

Collins, G. S., Dhiman, P., Navarro, C. L. A., Ma, J., Hooft, L., Reitsma, J. B., Logullo, P., Beam, A. L., Peng, L., Calster, B. V., Smeden, M. van, Riley, R. D., & Moons, K. G. (2021). Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*, *11*(7), e048008. https://doi.org/10.1136/bmjopen-2020-048008

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Calster, B. V., Ghassemi, M., Liu, X., Reitsma, J. B., Smeden, M. van, Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., … Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, *385*, e078378. https://doi.org/10.1136/bmj-2023-078378

Collins, G. S., Ogundimu, E. O., & Altman, D. G. (2016). Sample size considerations for the external validation of a multivariable prognostic model: A resampling study. *Statistics in Medicine*, *35*(2), 214–226. https://doi.org/10.1002/sim.6787

Constantino, M. J., Arnkoff, D. B., Glass, C. R., Ametrano, R. M., & Smith, J. Z. (2011). Expectations. *Journal of Clinical Psychology*, *67*(2), 184–192. https://doi.org/10.1002/jclp.20754

Cowlishaw, S., Metcalf, O., Stone, C., O’Donnell, M., Lotzin, A., Forbes, D., Hegarty, K., & Kessler, D. (2021). Posttraumatic Stress Disorder in Primary Care: A Study of General Practices in England. *Journal of Clinical Psychology in Medical Settings*, *28*(3), 427–435. https://doi.org/10.1007/s10880-020-09732-6

Creamer, M., Bell, R., & Failla, S. (2003). Psychometric properties of the Impact of Event Scale—Revised. *Behaviour Research and Therapy*, *41*(12), 1489–1496. https://doi.org/10.1016/j.brat.2003.07.010

Cronbach, R., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. Irvington. https://psycnet.apa.org/record/1978-11462-000

Da Costa, J. M. (1871). On irritable heart: A clinical study of a form of cardiac disorder and its consequences. *American Journal of the Clinical Sciences*, *61*, 17–52.

D’Agostino, R. B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, *17*(19), 2265–2281.

Dalmaijer, E. S., Nord, C. L., & Astle, D. E. (2022). Statistical power for cluster analysis. *BMC Bioinformatics*, *23*(1), 205. https://doi.org/10.1186/s12859-022-04675-1

de Jongh, A., de Roos, C., & El-Leithy, S. (2024). State of the science: Eye movement desensitization and reprocessing (EMDR) therapy. *Journal of Traumatic Stress*, *37*(2), 205–216. https://doi.org/10.1002/jts.23012

De Jongh, A., Resick, P. A., Zoellner, L. A., van Minnen, A., Lee, C. W., Monson, C. M., Foa, E. B., Wheeler, K., Broeke, E. ten, Feeny, N., Rauch, S. A. M., Chard, K. M., Mueser, K. T., Sloan, D. M., van der Gaag, M., Rothbaum, B. O., Neuner, F., de Roos, C., Hehenkamp, L. M. J., … Bicanic, I. A. E. (2016). Critical Analysis of the Current Treatment

Guidelines for Complex Ptsd in Adults. *Depression and Anxiety*, *33*(5), 359–369. https://doi.org/10.1002/da.22469

Debell, F., Fear, N. T., Head, M., Batt-Rawden, S., Greenberg, N., Wessely, S., & Goodwin, L. (2014). A systematic review of the comorbidity between PTSD and alcohol misuse. *Social Psychiatry and Psychiatric Epidemiology*, *49*(9), 1401–1425. https://doi.org/doi.org/10.1007/s00127-014-0855-7

Deisenhofer, A.-K., Barkham, M., Beierl, E. T., Schwartz, B., Aafjes-van Doorn, K., Beevers, C. G., Berwian, I. M., Blackwell, S. E., Bockting, C. L., Brakemeier, E.-L., Brown, G., Buckman, J. E. J., Castonguay, L. G., Cusack, C. E., Dalgleish, T., de Jong, K., Delgadillo, J., DeRubeis, R. J., Driessen, E., … Cohen, Z. D. (2024). Implementing precision methods in personalizing psychological therapies: Barriers and possible ways forward. *Behaviour Research and Therapy*, *172*, 104443. https://doi.org/10.1016/j.brat.2023.104443

Deisenhofer, A.-K., Delgadillo, J., Rubel, J. A., Bohnke, J. R., Zimmermann, D., Schwartz, B., & Lutz, W. (2018). Individual treatment selection for patients with posttraumatic stress disorder. *Depression and Anxiety*, *35*(6), 541–550. https://doi.org/10.1002/da.22755

Delamain, H., Buckman, J. E. J., O'Driscoll, C., Suh, J. W., Stott, J., Singh, S., Naqvi, S. A., Leibowitz, J., Pilling, S., & Saunders, R. (2024). Predicting post-treatment symptom severity for adults receiving psychological therapy in routine care for generalised anxiety disorder: A machine learning approach. *Psychiatry Research*, *336*, 115910. https://doi.org/10.1016/j.psychres.2024.115910

Delgadillo, J., Ali, S., Fleck, K., Agnew, C., Southgate, A., Parkhouse, L., Cohen, Z. D., DeRubeis, R. J., & Barkham, M. (2022). Stratified Care vs Stepped Care for Depression: A Cluster Randomized Clinical Trial. *JAMA Psychiatry*, *79*(2), 101–108. https://doi.org/10.1001/jamapsychiatry.2021.3539

Delgadillo, J., & Atzil-Slonim, D. (2022). Artificial intelligence, machine learning and mental health. In *Reference Module in Neuroscience and Biobehavioral Psychology*. Elsevier. https://eprints.whiterose.ac.uk/197827/

Delgadillo, J., & Gonzalez Salas Duhne, P. (2020). Targeted prescription of cognitive–behavioral therapy versus person-centered counseling for depression using a machine learning approach. *Journal of Consulting and Clinical Psychology*, *88*(1), 14–24. https://doi.org/10.1037/ccp0000476

Delgadillo, J., Huey, D., Bennett, H., & McMillan, D. (2017). Case complexity as a guide for psychological treatment selection. *Journal of Consulting and Clinical Psychology*, *85*(9), 835–853. https://doi.org/10.1037/ccp0000231

Delgadillo, J., & Lutz, W. (2020). A Development Pathway Towards Precision Mental Health Care. *JAMA Psychiatry*, *77*(9), 889–890. https://doi.org/10.1001/jamapsychiatry.2020.1048

Delgadillo, J., & Richardson, T. (2024). *On poverty and trauma: Associations between neighbourhood socioeconomic deprivation with post-traumatic stress disorder severity and treatment response*. [Manuscript in preparation].

Delgadillo, J., Rubel, J., & Barkham, M. (2020). Towards personalized allocation of patients to therapists. *Journal of Consulting and Clinical Psychology*, *88*(9), 799–808. https://doi.org/10.1037/ccp0000507

Department for Communities and Local Government. (2015). *The English indices of deprivation 2015*. Department for Communities and Local Government.

Department of Veterans Affairs and Department of Defense (VA/DoD). (2017). *VA/DoD clinical practice guideline for the management of posttraumatic stress disorder and acute stress disorder*. Author. https://www.healthquality.va.gov/guidelines/MH/ptsd/VADoDPTSDCPGFinal012418.pdf

Department of Veterans Affairs and Department of Defense (VA/DoD). (2023). *VA/DoD clinical practice guideline for the management of posttraumatic stress disorder and acute stress disorder (Version 4.0)*. Author. https://www.healthquality.va.gov/guidelines/MH/ptsd/VADoDPTSDCPGFinal012418.pdf

DeRubeis, R. J. (2019). The history, current status, and possible future of precision mental health. *Behaviour Research and Therapy*, *123*, 103506. https://doi.org/10.1016/j.brat.2019.103506

DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L. A., & Lorenzo-Luaces, L. (2014). The Personalized Advantage Index: Translating Research on Prediction into Individualized Treatment Recommendations. A Demonstration. *PLOS ONE*, *9*(1), e83875. https://doi.org/10.1371/journal.pone.0083875

Dewar, M., Paradis, A., & Fortin, C. A. (2020). Identifying Trajectories and Predictors of Response to Psychotherapy for Post-Traumatic Stress Disorder in Adults: A Systematic Review of Literature. *The Canadian Journal of Psychiatry*, *65*(2), 71–86. https://doi.org/10.1177/0706743719875602

Dumovich, J., & Singh, P. (2024). Physiology, Trauma. In *StatPearls*. StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK538478/

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine Learning Approaches for Clinical Psychology and Psychiatry. *Annual Review of Clinical Psychology*, *14*, 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Efron, B. (1983). Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *Journal of the American Statistical Association*, *78*(382), 316–331. https://doi.org/10.1080/01621459.1983.10477973

Ehlers, A., & Clark, D. M. (2000). A cognitive model of posttraumatic stress disorder. *Behaviour Research and Therapy*, *38*(4), 319–345. https://doi.org/10.1016/S0005-7967(99)00123-0

Ehlers, A., Clark, D. M., Hackmann, A., McManus, F., & Fennell, M. (2005). Cognitive therapy for post-traumatic stress disorder: Development and evaluation. *Behaviour Research and Therapy*, *43*(4), 413–431. https://doi.org/10.1016/j.brat.2004.03.006

Ehlers, A., & Wild, J. (2020). Cognitive therapy for PTSD. In *Casebook to the APA Clinical Practice Guideline for the treatment of PTSD* (pp. 91–121). American Psychological Association. https://doi.org/10.1037/0000196-005

Ehlers, A., & Wild, J. (2022). Cognitive Therapy for PTSD: Updating Memories and Meanings of Trauma. In U. Schnyder & M. Cloitre (Eds.), *Evidence Based Treatments for Trauma-Related Psychological Disorders: A Practical Guide for Clinicians* (pp. 181–210). Springer International Publishing. https://doi.org/10.1007/978-3-030-97802-0_9

Etkin, A., Maron-Katz, A., Wu, W., Fonzo, G. A., Huemer, J., Vertes, P. E., Patenaude, B., Richiardi, J., Goodkind, M. S., Keller, C. J., Ramos-Cejudo, J., Zaiko, Y. V., Peng, K. K., Shpigel, E., Longwell, P., Toll, R. T., Thompson, A., Zack, S., Gonzalez, B., … O'Hara, R. (2019). Using fMRI connectivity to define a treatment-resistant form of post-traumatic stress disorder. *Science Translational Medicine*, *11*(486), 1–12. https://doi.org/10.1126/scitranslmed.aal3236

Eulenburg, A. (1878). *Lehrbuch der Nervenkrankheite* (Vols. 1–2). August Hirschwald.

Fear, N. T., Bridges, S., Hatch, S., Hawkins, V., & Wessely, S. (2016). Chapter 4: Posttraumatic stress disorder. In McManus S, Bebbington P, Jenkins R, & Brugha T (Eds.), *Mental health and wellbeing in England: Adult Psychiatric Morbidity Survey 2014*. NHS Digital.

Firth, N., Saxon, D., Stiles, W. B., & Barkham, M. (2019). Therapist and clinic effects in psychotherapy: A three-level model of outcome variability. *Journal of Consulting and Clinical Psychology*, *87*(4), 345–356. https://doi.org/10.1037/ccp0000388

Fleming, C. E., Kholodkov, T., Dillon, K. H., Belvet, B., & Crawford, E. F. (2018). Actuarial prediction of psychotherapy retention among Iraq-Afghanistan veterans with posttraumatic stress disorder. *Psychol Serv*, *15*(2), 172–180. https://doi.org/10.1037/ser0000139

Foa, E. B., & Cahill, S. P. (2001). Psychological Therapies: Emotional Processing. In N. J. Smelser & P. B. Baltes (Eds.), *International Encyclopedia of the Social and Behavioral Sciences* (pp. 12363–12369). Elsevier.

Foa, E. B., Hembree, E. A., Rothbaum, B. O., & Rauch, S. A. M. (2019). *Prolonged exposure therapy for PTSD: Emotional processing of traumatic experience: Therapist guide* (2nd ed.). Oxford University Press. https://doi.org/10.1093/med-psych/9780190926939.001.0001

Foa, E. B., & Kozak, M. J. (1986). Emotional processing of fear: Exposure to corrective information. *Psychological Bulletin*, *99*(1), 20–35. https://doi.org/10.1037/0033-2909.99.1.20

Foa, E. B., Steketee, G., & Rothbaum, B. O. (1989). Behavioral/cognitive conceptualizations of post-traumatic stress disorder. *Behavior Therapy*, *20*(2), 155–176. https://doi.org/10.1016/S0005-7894(89)80067-X

Forbes, D., Creamer, M., Allen, N., Elliott, P., McHugh, T., Debenham, P., & Hopwood, M. (2003). MMPI-2 Based Subgroups of Veterans with Combat-related PTSD: Differential Patterns of Symptom Change After Treatment. *The Journal of Nervous and Mental Disease*, *191*(8), 531–537. https://doi.org/10.1097/01.nmd.0000082181.79051.83

Foresee, D. F., & Hagan, M. T. (1997). Gauss-Newton approximation to Bayesian learning. *Proceedings of International Conference on Neural Networks (ICNN'97)*, *3*, 1930–1935 vol.3. https://doi.org/10.1109/ICNN.1997.614194

Freud, S. (1962). The aetiology of hysteria (1896). In *Complete Psychological Works* (Vol. 3). Hogarth Press.

Galatzer-Levy, I. R., & Bryant, R. A. (2013). 636,120 Ways to Have Posttraumatic Stress Disorder. *Perspectives on Psychological Science*, *8*(6), 651–662. https://doi.org/10.1177/1745691613504115

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*(4), 1360–1383. https://doi.org/10.1214/08-AOAS191

Giesemann, J., Delgadillo, J., Schwartz, B., Bennemann, B., & Lutz, W. (2023). Predicting dropout from psychological treatment using different machine learning algorithms, resampling methods, and sample sizes. *Psychotherapy Research*, *33*(6), 683–695. https://doi.org/10.1080/10503307.2022.2161432

Glaz, A. L., Haralambous, Y., Kim-Dufor, D.-H., Lenca, P., Billot, R., Ryan, T. C., Marsh, J., DeVylder, J., Walter, M., Berrouiguet, S., & Lemey, C. (2021). Machine Learning and Natural Language Processing in Mental Health: Systematic Review. *Journal of Medical Internet Research*, *23*(5), e15708. https://doi.org/10.2196/15708

Gómez Penedo, J. M., Rubel, J., Meglio, M., Bornhauser, L., Krieger, T., Babl, A., Muiños, R., Roussos, A., Delgadillo, J., Flückiger, C., Berger, T., Lutz, W., & grosse Holtforth, M. (2023). Using machine learning algorithms to predict the effects of change processes in psychotherapy: Toward process-level treatment personalization. *Psychotherapy*, *60*(4), 536–547. https://doi.org/10.1037/pst0000507

Haddaway, N. R. (2021). *citationchaser: An R package and Shiny app for forward and backward citations chasing in academic searching*. https://estech.shinyapps.io/citationchaser/

Hahn, T., Nierenberg, A. A., & Whitfield-Gabrieli, S. (2017). Predictive analytics in mental health: Applications, guidelines, challenges and perspectives. *Molecular Psychiatry*, *22*(1), 37–43. https://doi.org/10.1038/mp.2016.201

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://hastie.su.domains/Papers/ESLII.pdf

Health Education England. (2021). *National Curriculum for Eye Movement Desensitisation and Reprocessing (EMDR) with Adults*. https://www.hee.nhs.uk/sites/default/files/documents/National%20Curriculum%20for%20EMDR%20Training%20Final%20-%202021.pdf

Hehlmann, M. I., Schwartz, B., Moggia, D., Schaffrath, J., & Lutz, W. (2023). *Cross-sample Validation of Affect and Rumination as Predictors of Outcome in Psychological Therapy [Manuscript in preparation].*

Held, P., Schubert, R. A., Pridgen, S., Kovacevic, M., Montes, M., Christ, N. M., Banerjee, U., & Smith, D. L. (2022). Who will respond to intensive PTSD treatment? A machine learning approach to predicting response prior to starting treatment. *Journal of Psychiatric Research*, *151*, 78–85. https://doi.org/10.1016/j.jpsychires.2022.03.066

Hendriks, L., De Kleine, R. A., Broekman, T. G., Hendriks, G.-J., & Van Minnen, A. (2018). Intensive prolonged exposure therapy for chronic PTSD patients following multiple trauma and multiple treatment attempts. *European Journal of Psychotraumatology*, *9*(1), 1425574. https://doi.org/10.1080/20008198.2018.1425574

Herman, J. L. (1992). Complex PTSD: A syndrome in survivors of prolonged and repeated trauma. *Journal of Traumatic Stress*, *5*(3), 377–391. https://doi.org/10.1002/jts.2490050305

Herzog, P., & Kaiser, T. (2022). Is it worth it to personalize the treatment of PTSD? – A variance-ratio meta-analysis and estimation of treatment effect heterogeneity in RCTs of PTSD. *Journal of Anxiety Disorders*, *91*, 102611. https://doi.org/10.1016/j.janxdis.2022.102611

Herzog, P., Voderholzer, U., Gärtner, T., Osen, B., Svitak, M., Doerr, R., Rolvering-Dijkstra, M., Feldmann, M., Rief, W., & Brakemeier, E. L. (2021). Predictors of outcome during inpatient psychotherapy for posttraumatic stress disorder: A single-treatment, multi-site, practice-based study. *Psychother Res*, *31*(4), 468–482. https://doi.org/10.1080/10503307.2020.1802081

Hines, L. A., Sundin, J., Rona, R. J., Wessely, S., & Fear, N. T. (2014). Posttraumatic Stress Disorder Post Iraq and Afghanistan: Prevalence among Military Subgroups. *The Canadian Journal of Psychiatry*, *59*(9), 468–479. https://doi.org/10.1177/070674371405900903

Hoeboer, C. M., de Kleine, R., Oprel, D., Schoorl, M., van der Does, W., & van Minnen, A. (2021). Does complex PTSD predict or moderate treatment outcomes of three variants of exposure therapy? *Journal of Anxiety Disorders*, *80*. https://doi.org/10.1016/j.janxdis.2021.102388

Hoeboer, C. M., Oprel, D. A. C., De Kleine, R. A., Schwartz, B., Deisenhofer, A. K., Schoorl, M., Van Der Does, W. A. J., van Minnen, A., & Lutz, W. (2021). Personalization of treatment for patients with childhoodabuse-related posttraumatic stress disorder. *Journal of Clinical Medicine*, *10*(19), 4522. https://doi.org/10.3390/jcm10194522

Hofner, B., Mayr, A., Robinzonov, N., & Schmid, M. (2014). Model-based boosting in R: A hands-on tutorial using the R package mboost. *Computational Statistics*, *29*(1), 3–35. https://doi.org/10.1007/s00180-012-0382-5

Hong, S., & Lynn, H. S. (2020). Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. *BMC Medical Research Methodology*, *20*(1), 199. https://doi.org/10.1186/s12874-020-01080-1

Hool, N. (2010). *BABCP Core Curriculum Reference Document*. BABCP. https://www.babcp.com/Portals/0/Files/About/BABCP-Core-Curriculum.pdf?ver=2019-03-11-141836-733

Horowitz, M. J. (1976). *Stress response syndromes* (pp. xvii, 366). Jason Aronson.

Huibers, M. J. H., Cohen, Z. D., Lemmens, L. H. J. M., Arntz, A., Peeters, F. P. M. L., Cuijpers, P., & DeRubeis, R. J. (2015). Predicting Optimal Outcomes in Cognitive Therapy or Interpersonal Psychotherapy for Depressed Individuals Using the Personalized Advantage Index Approach. *PLOS ONE*, *10*(11), e0140771. https://doi.org/10.1371/journal.pone.0140771

Iba, K., Shinozaki, T., Maruo, K., & Noma, H. (2021). Re-evaluation of the comparative effectiveness of bootstrap-based optimism correction methods in the development of multivariable clinical prediction models. *BMC Medical Research Methodology*, *21*(1), 9. https://doi.org/10.1186/s12874-020-01201-w

International Society for Traumatic Stress Studies (ISTSS). (2018). *ISTSS PTSD prevention and treatment guidelines: Methodology and recommendations*. http://www.istss.org/getattachment/TreatingTrauma/New-ISTSS-Prevention-and-Treatment-Guidelines/ISTSS_ PreventionTreatmentGuidelines_FNL-March-19-2019.pdf.aspx

Jacobson, N. S., & Truax, P. (1991). Clinical Significance: A Statistical Approach to Denning Meaningful Change in Psychotherapy Research. *Journal of Consulting and Clinical Psychology*, *59*(1), 12–19. https://doi.org/10.1037/10109-042

Jericho, B., Luo, A., & Berle, D. (2021). Trauma-focused psychotherapies for post-traumatic stress disorder: A systematic review and network meta-analysis. *Acta Psychiatrica Scandinavica*. https://doi.org/10.1111/acps.13366

Jones, S. (2017). Describing the Mental Health Profile of First Responders: A Systematic Review. *Journal of the American Psychiatric Nurses Association*, *23*(3), 200–214. https://doi.org/10.1177/1078390317695266

Karatzias, T., Hyland, P., Bradley, A., Cloitre, M., Roberts, N. P., Bisson, J. I., & Shevlin, M. (2019). Risk factors and comorbidity of ICD-11 PTSD and complex PTSD: Findings from a trauma-exposed population based sample of adults in the United Kingdom. *Depression and Anxiety*, *36*(9), 887–894. https://doi.org/10.1002/da.22934

Kardiner, A. (1941). *The traumatic neuroses of war* (pp. x, 258). National Research Council. https://doi.org/10.1037/10581-000

Keefe, J. R., Stirman, S. W., Cohen, Z. D., DeRubeis, R. J., Smith, B. N., & Resick, P. A. (2018). In rape trauma PTSD, patient characteristics indicate which trauma-focused treatment they are most likely to complete. *Depression and Anxiety*, *35*(4), 330–338. https://doi.org/10.1002/da.22731

Kessler, R. C., Bossarte, R. M., Luedtke, A., Zaslavsky, A. M., & Zubizarreta, J. R. (2019). Machine learning methods for developing precision treatment rules with observational data. *Behaviour Research and Therapy*, *120*, 103412.

Kessler, R. C., Loo, H. M. van, Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Ebert, D. D., Jonge, P. de, Nierenberg, A. A., Rosellini, A. J., Sampson, N. A., Schoevers, R. A., Wilcox, M. A., & Zaslavsky, A. M. (2017). Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. *Epidemiology and Psychiatric Sciences*, *26*(1), 22–36. https://doi.org/10.1017/S2045796016000020

Kessler, R. C., & Luedtke, A. (2021). Pragmatic Precision Psychiatry—A New Direction for Optimizing Treatment Selection. *JAMA Psychiatry*, *78*(12), 1384–1390. https://doi.org/10.1001/jamapsychiatry.2021.2500

Kirkbride, J. B., Anglin, D. M., Colman, I., Dykxhoorn, J., Jones, P. B., Patalay, P., Pitman, A., Soneson, E., Steare, T., Wright, T., & Griffiths, S. L. (2024). The social determinants of mental health and disorder: Evidence, prevention and recommendations. *World Psychiatry*, *23*(1), 58–90. https://doi.org/10.1002/wps.21160

Knefel, M., Tran, U. S., & Lueger-Schuster, B. (2016). The association of posttraumatic stress disorder, complex posttraumatic stress disorder, and borderline personality disorder from a network analytical perspective. *Journal of Anxiety Disorders*, *43*, 70–78. https://doi.org/10.1016/j.janxdis.2016.09.002

Koenen, K. C., Ratanatharathorn, A., Ng, L., McLaughlin, K. A., Bromet, E. J., Stein, D. J., Karam, E. G., Ruscio, A. M., Benjet, C., Scott, K., Atwoli, L., Petukhova, M., Lim, C. C. W., Aguilar-Gaxiola, S., Al-Hamzawi, A., Alonso, J., Bunting, B., Ciutan, M., Girolamo, G. de, … Kessler, R. C. (2017). Posttraumatic stress disorder in the World Mental Health Surveys. *Psychological Medicine*, *47*(13), 2260–2274. https://doi.org/10.1017/S0033291717000708

Kraemer, H. C., & Kupfer, D. J. (2006). Size of treatment effects and their importance to clinical research and practice. *Biological Psychiatry*, *59*(11), 990–996. https://doi.org/10.1016/j.biopsych.2005.09.014

Kratzer, L., Heinz, P., Schennach, R., Schiepek, G. K., Padberg, F., & Jobst, A. (2019). Inpatient Treatment of Complex PTSD Following Childhood Abuse: Effectiveness and Predictors of Treatment Outcome. *PPmP Psychotherapie Psychosomatik Medizinische Psychologie*, *69*(3–4), 114–122. https://doi.org/10.1055/a-0591-3962

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Kruschke, J. K. (2011). Bayesian Assessment of Null Values Via Parameter Estimation and Model Comparison. *Perspectives on Psychological Science*, *6*(3), 299–312. https://doi.org/10.1177/1745691611406925

Kyriacou, D. N., & Lewis, R. J. (2016). Confounding by Indication in Clinical Research. *JAMA*, *316*(17), 1818–1819. https://doi.org/10.1001/jama.2016.16435

Lamprecht, F., & Sack, M. (2002). Posttraumatic Stress Disorder Revisited. *Psychosomatic Medicine*, *64*(2), 222.

https://journals.lww.com/psychosomaticmedicine/fulltext/2002/03000/Posttraumatic_Stress_Disorder_Revisited.5.aspx

Landin-Romero, R., Moreno-Alcazar, A., Pagani, M., & Amann, B. L. (2018). How Does Eye Movement Desensitization and Reprocessing Therapy Work? A Systematic Review on Suggested Mechanisms of Action. *Frontiers in Psychology*, *9*. https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01395

Lee, D. J., Schnitzlein, C. W., Wolf, J. P., Vythilingam, M., Rasmusson, A. M., & Hoge, C. W. (2016). Psychotherapy Versus Pharmacotherapy for Posttraumatic Stress Disorder: Systemic Review and Meta-Analyses to Determine First-Line Treatments. *Depression and Anxiety*, *33*(9), 792–806. https://doi.org/10.1002/da.22511

Lee, Y., Ragguett, R.-M., Mansur, R. B., Boutilier, J. J., Rosenblat, J. D., Trevizol, A., Brietzke, E., Lin, K., Pan, Z., Subramaniapillai, M., Chan, T. C. Y., Fus, D., Park, C., Musial, N., Zuckerman, H., Chen, V. C.-H., Ho, R., Rong, C., & McIntyre, R. S. (2018). Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. *Journal of Affective Disorders*, *241*, 519–532. https://doi.org/10.1016/j.jad.2018.08.073

Lenhard, W., & Lenhard, A. (2016). *Computation of effect sizes*. Psychometrica. https://www.psychometrica.de/effect_size.html

Lewis, C., Roberts, N. P., Andrew, M., Starling, E., & Bisson, J. I. (2020). Psychological therapies for post-traumatic stress disorder in adults: Systematic review and meta-analysis. *European Journal of Psychotraumatology*, *11*(1), 1729633. https://doi.org/10.1080/20008198.2020.1729633

Lewis, C., Roberts, N. P., Gibson, S., & Bisson, J. I. (2020). Dropout from psychological therapies for post-traumatic stress disorder (PTSD) in adults: Systematic review and meta-analysis. *European Journal of Psychotraumatology*, *11*(1), 1709709. https://doi.org/10.1080/20008198.2019.1709709

López-Castro, T., Zhao, Y., Fitzpatrick, S., Ruglass, L. M., & Hien, D. A. (2021). Seeing the forest for the trees: Predicting attendance in trials for co-occurring PTSD and substance use disorders with a machine learning approach. *Journal of Consulting and Clinical Psychology*, *89*(10), 869–884. https://doi.org/10.1037/ccp0000688

Lorimer, B., Kellett, S., Giesemann, J., Lutz, W., & Delgadillo, J. (2024). An investigation of treatment return after psychological therapy for depression and anxiety. *Behavioural and Cognitive Psychotherapy*, *52*(2), 149–162. https://doi.org/10.1017/S1352465823000322

Lowe, S. R., Galea, S., Uddin, M., & Koenen, K. C. (2014). Trajectories of Posttraumatic Stress Among Urban Residents. *American Journal of Community Psychology*, *53*(1–2), 159–172. https://doi.org/10.1007/s10464-014-9634-6

Luborsky, L., Singer, B., & Luborsky, L. (1975). Comparative Studies of Psychotherapies: Is It True That "Everyone Has Won and All Must Have Prizes"? *Archives of General Psychiatry*, *32*(8), 995–1008. https://doi.org/10.1001/archpsyc.1975.01760260059004

Luedtke, A., Sadikova, E., & Kessler, R. C. (2019). Sample size requirements for multivariate models to predict between-patient differences in best treatments of major depressive disorder. *Clinical Psychological Science*, *7*(3), 445–461. https://doi.org/10.1177/2167702618815466

Lutz, W., Deisenhofer, A.-K., Rubel, J., Bennemann, B., Giesemann, J., Poster, K., & Schwartz, B. (2022). Prospective evaluation of a clinical decision support system in psychological therapy. *Journal of Consulting and Clinical Psychology*, *90*(1), 90–106. https://doi.org/10.1037/ccp0000642

Lutz, W., Rubel, J. A., Schwartz, B., Schilling, V., & Deisenhofer, A.-K. (2019). Towards integrating personalized feedback research into clinical practice: Development of the Trier Treatment Navigator (TTN). *Behaviour Research and Therapy*, *120*, 103438. https://doi.org/10.1016/j.brat.2019.103438

Maercker, A. (2021). Development of the new CPTSD diagnosis for ICD-11. *Borderline Personality Disorder and Emotion Dysregulation*, *8*(1), 7. https://doi.org/10.1186/s40479-021-00148-8

Malejko, K., Abler, B., Plener, P. L., & Straub, J. (2017). Neural Correlates of Psychotherapeutic Treatment of Post-traumatic Stress Disorder: A Systematic Literature Review. *Frontiers in Psychiatry*, *8*, 85. https://doi.org/10.3389/fpsyt.2017.00085

Malgaroli, M., & Schultebraucks, K. (2021). Artificial Intelligence and Posttraumatic Stress Disorder (PTSD). *European Psychologist*, *25*(4), 272–282. https://doi.org/10.1027/1016-9040/a000423

Manchia, M., Pisanu, C., Squassina, A., & Carpiniello, B. (2020). Challenges and Future Prospects of Precision Medicine in Psychiatry. *Pharmacogenomics and Personalized Medicine*, *13*, 127–140. https://www.tandfonline.com/doi/full/10.2147/PGPM.S198225

Martin, A., Naunton, M., Kosari, S., Peterson, G., Thomas, J., & Christenson, J. K. (2021). Treatment Guidelines for PTSD: A Systematic Review. *Journal of Clinical Medicine*, *10*(18), 4175. https://doi.org/10.3390/jcm10184175

Mavranezouli, I., Megnin-Viggars, O., Daly, C., Dias, S., Welton, N. J., Stockton, S., Bhutani, G., Grey, N., Leach, J., Greenberg, N., Katona, C., El-Leithy, S., & Pilling, S. (2020). Psychological treatments for post-traumatic stress disorder in adults: A network meta-analysis. *Psychological Medicine*, *50*(4), 542–555. https://doi.org/10.1017/S0033291720000070

McLean, C. P., & Foa, E. B. (2024). State of the Science: Prolonged exposure therapy for the treatment of posttraumatic stress disorder. *Journal of Traumatic Stress*, *37*(4), 535–550. https://doi.org/10.1002/jts.23046

McNally, R. J. (2003). Progress and controversy in the study of posttraumatic stress disorder. *Annual Review of Psychology*, *54*, 229–252. https://doi.org/10.1146/annurev.psych.54.101601.145112

Meehan, A. J., Lewis, S. J., Fazel, S., Fusar-Poli, P., Steyerberg, E. W., Stahl, D., & Danese, A. (2022). Clinical prediction models in psychiatry: A systematic review of two decades of progress and challenges. *Molecular Psychiatry*, *27*(6), Article 6. https://doi.org/10.1038/s41380-022-01528-4

Meehl, P. E. (1954). *Clinical versus statistical prediction*. University of Minnesota Press.

Merz, J., Schwarzer, G., & Gerger, H. (2019). Comparative Efficacy and Acceptability of Pharmacological, Psychotherapeutic, and Combination Treatments in Adults with Posttraumatic Stress Disorder: A Network Meta-analysis. *JAMA Psychiatry*, *76*(9), 904–913. https://doi.org/10.1001/jamapsychiatry.2019.0951

Minami, T., Serlin, R. C., Wampold, B. E., Kircher, J. C., & Brown, G. (2008). Using clinical trials to benchmark effects produced in clinical practice. *Quality and Quantity*, *42*, 513–525.

Mitchell, M. (1998). *An introduction to genetic algorithms*. MIT press. https://link.springer.com/article/10.1007/s11135-006-9057-z

Moggia, D., Saxon, D., Lutz, W., Hardy, G. E., & Barkham, M. (2023). Applying precision methods to treatment selection for moderate/severe depression in person-centered experiential therapy or cognitive behavioral therapy. *Psychotherapy Research*, 1035-1050. https://doi.org/10.1080/10503307.2023.2269297

Moons, K. G. M., de Groot, J. A. H., Bouwmeester, W., Vergouwe, Y., Mallett, S., Altman, D. G., Reitsma, J. B., & Collins, G. S. (2014). Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. *PLoS Medicine*, *11*(10), e1001744. https://doi.org/10.1371/journal.pmed.1001744

Moons, K. G. M., Wolff, R. F., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. *Annals of Internal Medicine*, *170*(1), 1–33. https://doi.org/10.7326/M18-1377

Moskowitz, A., Dorahy, M. J., & Schäfer, I. (2019). *Psychosis, Trauma and Dissociation: Evolving Perspectives on Severe Psychopathology*. John Wiley & Sons, Incorporated. https://onlinelibrary.wiley.com/doi/book/10.1002/9781118585948

Mott, F. W. (1919). *War neuroses and shell shock*. Oxford University Press.

Mundt, J. C., Marks, I. M., Shear, M. K., & Greist, J. M. (2002). The Work and Social Adjustment Scale: A simple measure of impairment in functioning. *The British Journal of Psychiatry*, *180*(5), 461–464. https://doi.org/10.1192/bjp.180.5.461

Murray, H., Grey, N., Warnock-Parkes, E., Kerr, A., Wild, J., Clark, D. M., & Ehlers, A. (2022). Ten misconceptions about trauma-focused CBT for PTSD. *The Cognitive Behaviour Therapist*, *15*, e33. https://doi.org/10.1017/S1754470X22000307

Mushagalusa, C. A., Fandohan, A. B., & Glèlè Kakaï, R. (2022). Random Forests in Count Data Modelling: An Analysis of the Influence of Data Features and Overdispersion on Regression Performance. *Journal of Probability and Statistics*, *2022*, 1–21. https://doi.org/10.1155/2022/2833537

Myers, A. B. R. (1870). *On the etiology and prevalence of diseases of the heart among soldiers*. J. Churchill.

Myers, A. B. R. (1915). A contribution to the study of shell shock. *Lancet*, *188*, 316–320.

Najavits, L. M. (2015). The problem of dropout from "gold standard" PTSD therapies. *F1000Prime Reports*, *7*, 43. https://doi.org/10.12703/P7-43

National Collaborating Centre for Mental Health. (2018). *The NHS Talking Therapies manual*. https://www.england.nhs.uk/publication/the-improving-access-to-psychological-therapies-manual/

National Institute for Health and Care Excellence [NICE]. (2018a). *Evidence review D: Psychological, psychosocial and other non-pharmacological interventions for the treatment of PTSD in adults (NICE guideline [NG116])*. NICE. https://www.nice.org.uk/guidance/ng116/evidence

National Institute for Health and Care Excellence [NICE]. (2018b). *Post-traumatic stress disorder (NICE guideline [NG116])*. NICE. https://www.nice.org.uk/guidance/ng116

Ng, W., Minasny, B., Mendes, W. de S., & Demattê, J. A. M. (2020). The influence of training sample size on the accuracy of deep learning models for the prediction of soil properties with near-infrared spectroscopy data. *SOIL*, *6*(2), 565–578. https://doi.org/10.5194/soil-6-565-2020

NHS Digital. (2021, January). *Announcement of methodological change Improving Access to Psychological Therapies (IAPT) reports*. NHS England Digital. https://digital.nhs.uk/data-and-information/find-data-and-publications/statement-of-administrative-sources/methodological-changes/improving-access-to-psychological-therapies-iapt-reports-january-2021

NHS Digital. (2024). *NHS Talking Therapies, for anxiety and depression, Annual reports, 2022-23*. NHS Talking Therapies, for Anxiety and Depression, Annual Reports. https://digital.nhs.uk/data-and-information/publications/statistical/nhs-talking-therapies-for-anxiety-and-depression-annual-reports/2022-23

NHS England. (2014). *Improving Access to Psychological Therapies: Measuring Improvement and Recovery, Adult Services, Version 2*. http://www.oxfordahsn.org/wp-content/uploads/2015/11/measuring-recovery-2014.pdf

Nixon, R., King, M., Smith, B., Gradus, J., Resick, P., & Galovski, T. (2021). Predicting response to cognitive processing therapy for PTSD: a machine-learning approach. *Behaviour Research and Therapy*, *144*, 103920. https://doi.org/10.1016/j.brat.2021.103920

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M. K., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology*, *73*, 719–748.

Nye, A., Delgadillo, J., & Barkham, M. (2023). Efficacy of personalized psychological interventions: A systematic review and meta-analysis. *Journal of Consulting and Clinical Psychology*, *91*(7), 389-397. https://doi.org/10.1037/ccp0000820

O’Donovan, A., Cohen, B. E., Seal, K. H., Bertenthal, D., Margaretten, M., Nishimi, K., & Neylan, T. C. (2015). Elevated Risk for Autoimmune Disorders in Iraq and Afghanistan Veterans with Posttraumatic Stress Disorder. *Biological Psychiatry*, *77*(4), 365–374. https://doi.org/10.1016/j.biopsych.2014.06.015

Office for National Statistics. (n.d.). *Ethnic group, national identity and religion*. Retrieved May 17, 2023, from https://www.ons.gov.uk/methodology/classificationsandstandards/measuringequality/ethnicgroupnationalidentityandreligion

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Oppenheim, H. (1889). *Die traumatischen Neurosen*. A. Hirschwald.

Oxford University Press. (2023). Trauma. In *Oxford English Dictionary*. https://doi.org/10.1093/OED/6758286467

Pacella, M. L., Hruska, B., & Delahanty, D. L. (2013). The physical health consequences of PTSD and PTSD symptoms: A meta-analytic review. *Journal of Anxiety Disorders*, *27*(1), 33–46. https://doi.org/10.1016/j.janxdis.2012.08.004

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Pai, A., Suris, A. M., & North, C. S. (2017). Posttraumatic Stress Disorder in the DSM-5: Controversy, Change, and Conceptual Considerations. *Behavioral Sciences*, *7*(1), Article 1. https://doi.org/10.3390/bs7010007

Palazón-Bru, A., Martín-Pérez, F., Mares-García, E., Beneyto-Ripoll, C., Gil-Guillén, V. F., Pérez-Sempere, Á., & Carbonell-Torregrosa, M. Á. (2020). A general presentation on how to carry out a CHARMS analysis for prognostic multivariate models. *Statistics in Medicine*, *39*(23), 3207–3225. https://doi.org/10.1002/sim.8660

Paul, G. L. (1967). Strategy of outcome research in psychotherapy. *Journal of Consulting Psychology*, *31*(2), 109–118. https://doi.org/10.1037/h0024436

Persons, J. B. (2022). Case Formulation. *Cognitive and Behavioral Practice*, *29*(3), 537–540. https://doi.org/10.1016/j.cbpra.2022.02.014

Phoenix Australia Centre for Posttraumatic Mental Health. (2021). *Australian Guidelines for the Prevention and Treatment of Acute Stress Disorder, Posttraumatic Stress Disorder and Complex PTSD*. Author. https://www.phoenixaustralia.org/australian-guidelines-for-ptsd/

Ramchand, R., Rudavsky, R., Grant, S., Tanielian, T., & Jaycox, L. (2015). Prevalence of, Risk Factors for, and Consequences of Posttraumatic Stress Disorder and Other Mental Health Problems in Military Populations Deployed to Iraq and Afghanistan. *Current Psychiatry Reports*, *17*(5), 37. https://doi.org/10.1007/s11920-015-0575-z

Ramos-Lima, L. F., Waikamp, V., Antonelli-Salgado, T., Passos, I. C., & Freitas, L. H. M. (2020). The use of machine learning techniques in trauma-related disorders: A systematic review. *Journal of Psychiatric Research*, *121*, 159–172. https://doi.org/10.1016/j.jpsychires.2019.12.001

Ravi, M., Powers, A., Rothbaum, B. O., Stevens, J. S., & Michopoulos, V. (2023). Neighborhood poverty prospectively predicts PTSD symptoms six-months following trauma exposure. *Mental Health Science*, *1*(4), 213–221. https://doi.org/10.1002/mhs2.35

Ray, S. L. (2008). Evolution of Posttraumatic Stress Disorder and Future Directions. *Archives of Psychiatric Nursing*, *22*(4), 217–225. https://doi.org/10.1016/j.apnu.2007.08.005

Resick, P. A., Bovin, M. J., Calloway, A. L., Dick, A. M., King, M. W., Mitchell, K. S., Suvak, M. K., Wells, S. Y., Stirman, S. W., & Wolf, E. J. (2012). A critical evaluation of the complex PTSD literature: Implications for DSM-5. *Journal of Traumatic Stress*, *25*(3), 241–251. https://doi.org/10.1002/jts.21699

Resick, P. A., Galovski, T. E., Uhlmansiek, M. O., Scher, C. D., Clum, G. A., & Young-Xu, Y. (2008). A randomized clinical trial to dismantle components of cognitive processing therapy for posttraumatic stress disorder in female victims of interpersonal violence. *Journal of Consulting and Clinical Psychology*, *76*(2), 243–258. https://doi.org/10.1037/0022-006X.76.2.243

Resick, P. A., Monson, M. M., & Chard, K. M. (2017). *Cognitive Processing Therapy for PTSD: A Comprehensive Manual*. Guilford Press.

Resick, P. A., & Schnicke, M. K. (1992). Cognitive processing therapy for sexual assault victims. *Journal of Consulting and Clinical Psychology*, *60*(5), 748–756. https://doi.org/10.1037/0022-006X.60.5.748

Richards, D. A., & Borglin, G. (2011). Implementation of psychological therapies for anxiety and depression in routine practice: Two year prospective cohort study. *Journal of Affective Disorders*, *133*(1–2), 51–60. https://doi.org/10.1016/j.jad.2011.03.024

Riley, R. D., Ensor, J., Snell, K. I. E., Harrell, F. E., Martin, G. P., Reitsma, J. B., Moons, K. G. M., Collins, G., & Smeden, M. van. (2020). Calculating the sample size required for developing a clinical prediction model. *BMJ*, *368*, m441. https://doi.org/10.1136/bmj.m441

Riley, R. D., Snell, K. I. E., Martin, G. P., Whittle, R., Archer, L., Sperrin, M., & Collins, G. S. (2021). Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. *Journal of Clinical Epidemiology*, *132*, 88–96. https://doi.org/10.1016/j.jclinepi.2020.12.005

Robinson, L., Kellett, S., & Delgadillo, J. (2020). Dose-response patterns in low and high intensity cognitive behavioral therapy for common mental health problems. *Depression and Anxiety*, *37*(3), 285–294. https://doi.org/10.1002/da.22999

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rosenzweig, S. (1936). Some implicit common factors in diverse methods of psychotherapy. *American Journal of Orthopsychiatry*, *6*(3), 412–415. https://doi.org/10.1111/j.1939-0025.1936.tb05248.x

Sajjadian, M., Lam, R. W., Milev, R., Rotzinger, S., Frey, B. N., Soares, C. N., Parikh, S. V., Foster, J. A., Turecki, G., Müller, D. J., Strother, S. C., Farzan, F., Kennedy, S. H., & Uher, R. (2021). Machine learning in the prediction of depression treatment outcomes: A systematic review and meta-analysis. *Psychological Medicine*, *51*(16), 2742–2751. https://doi.org/10.1017/S0033291721003871

Salazar de Pablo, G., Studerus, E., Vaquerizo-Serrano, J., Irving, J., Catalan, A., Oliver, D., Baldwin, H., Danese, A., Fazel, S., Steyerberg, E. W., Stahl, D., & Fusar-Poli, P. (2021). Implementing Precision Psychiatry: A Systematic Review of Individualized Prediction Models for Clinical Practice. *Schizophrenia Bulletin*, *47*(2), 284–297. https://doi.org/10.1093/schbul/sbaa120

Sareen, J., Cox, B. J., Stein, M. B., Afifi, T. O., Fleet, C., & Asmundson, G. J. (2007). Physical and mental comorbidity, disability, and suicidal behavior associated with posttraumatic stress disorder in a large community sample. *Psychosomatic Medicine*, *69*(3), 242–248. https://doi.org/10.1097/PSY.0b013e31803146d8

Scheiderer, E. M., Wood, P. K., & Trull, T. J. (2015). The comorbidity of borderline personality disorder and posttraumatic stress disorder: Revisiting the prevalence and associations in a general population sample. *Borderline Personality Disorder and Emotion Dysregulation*, *2*(1), 11. https://doi.org/10.1186/s40479-015-0032-y

Schottenbauer, M. A., Glass, C. R., Arnkoff, D. B., Tendick, V., & Gray, S. H. (2008). Nonresponse and Dropout Rates in Outcome Studies on PTSD: Review and Methodological Considerations. *Psychiatry: Interpersonal and Biological Processes*, *71*(2), 134–168. https://doi.org/10.1521/psyc.2008.71.2.134

Schwartz, B., Cohen, Z. D., Rubel, J. A., Zimmermann, D., Wittmann, W. W., & Lutz, W. (2021). Personalized treatment selection in routine care: Integrating machine learning and statistical algorithms to recommend cognitive behavioral or psychodynamic therapy. *Psychotherapy Research*, *31*(1), 33–51. https://doi.org/10.1080/10503307.2020.1769219

Schwartz, R. A., Chambless, D. L., Barber, J. P., & Milrod, B. (2021). Testing clinical intuitions about barriers to improvement in cognitive-behavioral therapy for panic disorder. *Behavior Therapy*, *52*(4), 956–969. https://dx.doi.org/10.1016/j.beth.2020.12.004

Shalev, A., Liberzon, I., & Marmar, C. (2017). Post-Traumatic Stress Disorder. *New England Journal of Medicine*, *376*(25), 2459–2469. https://doi.org/10.1056/NEJMra1612499

Shapiro, F. (2001). *Eye movement desensitization and reprocessing (EMDR): Basic principles, protocols, and procedures*. Guilford Press.

Shapiro, F. (2018). *Eye movement desensitization and reprocessing (EMDR): Basic principles, protocols, and procedures* (3rd Edition). Guilford Press.

Shapiro, F., & Maxfield, L. (2002). Eye Movement Desensitization and Reprocessing. In M. Hersen & W. Sledge (Eds.), *Encyclopedia of Psychotherapy*. Elsevier Science.

Shekelle, P. G., Ortiz, E., Rhodes, S., Morton, S. C., Eccles, M. P., Grimshaw, J. M., & Woolf, S. H. (2001). Validity of the Agency for Healthcare Research and Quality Clinical Practice

GuidelinesHow Quickly Do Guidelines Become Outdated? *JAMA*, *286*(12), 1461–1467. https://doi.org/10.1001/jama.286.12.1461

Shevlin, M., McBride, O., Murphy, J., Miller, J. G., Hartman, T. K., Levita, L., Mason, L., Martinez, A. P., McKay, R., & Stocks, T. V. (2020). Anxiety, depression, traumatic stress and COVID-19-related anxiety in the UK general population during the COVID-19 pandemic. *BJPsych Open*, *6*(6) , e125. https://doi.org/10.1192/bjo.2020.109

Simiola, V., Neilson, E. C., Thompson, R., & Cook, J. M. (2015). Preferences for trauma treatment: A systematic review of the empirical literature. *Psychological Trauma: Theory, Research, Practice, and Policy*, *7*(6), 516–524. https://doi.org/10.1037/tra0000038

Simmonds-Buckley, M., Catarino, A., & Delgadillo, J. (2021). Depression subtypes and their response to cognitive behavioral therapy: A latent transition analysis. *Depression and Anxiety*, *38*(9), 907–916. https://doi.org/10.1002/da.23161

Siontis, G. C. M., Tzoulaki, I., Castaldi, P. J., & Ioannidis, J. P. A. (2015). External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *Journal of Clinical Epidemiology*, *68*(1), 25–34. https://doi.org/10.1016/j.jclinepi.2014.09.007

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, *32*(9), 752–760. https://doi.org/10.1037/0003-066X.32.9.752

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*(3), 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

Snow, R. E. (1991). Aptitude-treatment interaction as a framework for research on individual differences in psychotherapy. *Journal of Consulting and Clinical Psychology*, *59*(2), 205–216. https://doi.org/10.1037/0022-006X.59.2.205

Spinhoven, P., Penninx, B. W., van Hemert, A. M., de Rooij, M., & Elzinga, B. M. (2014). Comorbidity of PTSD in anxiety and depressive disorders: Prevalence and shared risk factors. *Child Abuse & Neglect*, *38*(8), 1320–1330. https://doi.org/10.1016/j.chiabu.2014.01.017

Spitzer, R. L., First, M. B., & Wakefield, J. C. (2007). Saving PTSD from itself in DSM-V. *Journal of Anxiety Disorders*, *21*(2), 233–241. https://doi.org/10.1016/j.janxdis.2006.09.006

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

Steenkamp, M. M., Litz, B. T., Hoge, C. W., & Marmar, C. R. (2015). Psychotherapy for Military-Related PTSD: A Review of Randomized Clinical Trials. *JAMA*, *314*(5), 489–500. https://doi.org/10.1001/jama.2015.8370

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Steyerberg, E. W. (2019). *Clinical prediction models: A practical approach to development, validation, and updating* (2nd ed.). Springer Nature. https://link.springer.com/book/10.1007/978-3-030-16399-0

Steyerberg, E. W., Bleeker, S. E., Moll, H. A., Grobbee, D. E., & Moons, K. G. M. (2003). Internal and external validation of predictive models: A simulation study of bias and precision in small samples. *Journal of Clinical Epidemiology*, *56*(5), 441–447. https://doi.org/10.1016/S0895-4356(03)00047-7

Steyerberg, E. W., Harrell, F. E., Borsboom, G. J. J. M., Eijkemans, M. J. C., Vergouwe, Y., & Habbema, J. D. F. (2001). Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology*, *54*(8), 774–781. https://doi.org/10.1016/S0895-4356(01)00341-9

Stirman, S., Cohen, Z., Lunney, C., DeRubeis, R., Wiley, J., & Schnurr, P. (2021). A personalized index to inform selection of a trauma-focused or non-trauma-focused treatment for PTSD. *Behaviour Research and Therapy*, *142*, 103872. https://doi.org/10.1016/j.brat.2021.103872

Stuart, E. A., King, G., Imai, K., & Ho, D. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8), 1–28. https://doi.org/10.18637/jss.v042.i08

Stuke, H., Schoofs, N., Johanssen, H., Bermpohl, F., Ülsmann, D., Schulte-Herbrüggen, O., & Priebe, K. (2021). Predicting outcome of daycare cognitive behavioural therapy in a naturalistic sample of patients with PTSD: a machine learning approach. *European Journal of Psychotraumatology*, *12*(1), 1958471. https://doi.org/10.1080/20008198.2021.1958471

Su, X., Yan, X., & Tsai, C.-L. (2012). Linear regression. *WIREs Computational Statistics*, *4*(3), 275–294. https://doi.org/10.1002/wics.1198

Swift, J. K., Greenberg, R. P., Tompkins, K. A., & Parkin, S. R. (2017). Treatment refusal and premature termination in psychotherapy, pharmacotherapy, and their combination: A meta-analysis of head-to-head comparisons. *Psychotherapy*, *54*(1), 47–57. https://doi.org/10.1037/pst0000104

Tait, J., Kellett, S., Saxon, D., Deisenhofer, A.-K., Lutz, W., Barkham, M., & Delgadillo, J. (2024). Individual treatment selection for patients with post-traumatic stress disorder: External validation of a personalised advantage index. *Psychotherapy Research*, 1–14. https://doi.org/10.1080/10503307.2024.2360449

Tantithamthavorn, C., McIntosh, S., Hassan, A. E., & Matsumoto, K. (2017). An Empirical Comparison of Model Validation Techniques for Defect Prediction Models. *IEEE Transactions on Software Engineering*, *43*(1), 1–18. https://doi.org/10.1109/TSE.2016.2584050

Thériault, R., Ben-Shachar, M. S., Patil, I., Lüdecke, D., Wiernik, B. M., & Makowski, D. (2024). Check your outliers! An introduction to identifying statistical outliers in R with easystats. *Behavior Research Methods*, *56*(4), 4162–4172. https://doi.org/10.3758/s13428-024-02356-w

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Turnbull, G. J. (1998). A review of post-traumatic stress disorder. Part I: Historical development and classification. *Injury*, *29*(2), 87–91. https://doi.org/10.1016/s0020-1383(97)00131-9

Van Bronswijk, S. C., Bruijniks, S. J., Lorenzo-Luaces, L., Derubeis, R. J., Lemmens, L. H., Peeters, F. P., & Huibers, M. J. (2021). Cross-trial prediction in psychotherapy: External validation of the Personalized Advantage Index using machine learning in two Dutch randomized trials comparing CBT versus IPT for depression. *Psychotherapy Research*, *31*(1), 78–91.

Van Der Hart, O., & Brown, P. (1990). Concept of psychological trauma. *The American Journal of Psychiatry*, *147*, 1691. https://doi.org/10.1176/ajp.147.12.1691a

Van der Kolk, B. A., & Van der Hart, O. (1989). Pierre Janet and the breakdown of adaptation in psychological trauma. *American Journal of Psychiatry*, *146*(12), 1530–1540. https://www.besselvanderkolk.com/uploads/docs/janet_am_j_psychiat.pdf

van der Ploeg, T., Austin, P. C., & Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology*, *14*(1), 137. https://doi.org/10.1186/1471-2288-14-137

Vapnik, V., Golowich, S. E., & Smola, A. J. (1995). Support Vector Method for Function Approximation, Regression Estimation and Signal Processing. *Advances in Neural Information Processing Systems*, *9*, 281–287.

https://proceedings.neurips.cc/paper/1996/file/4f284803bd0966cc24fa8683a34afc6e-Paper.pdf

Vieira, S., Liang, X., Guiomar, R., & Mechelli, A. (2022). Can we predict who will benefit from cognitive-behavioural therapy? A systematic review and meta-analysis of machine learning studies. *Clinical Psychology Review, 97*, 102193. https://doi.org/10.1016/j.cpr.2022.102193

Voorendonk, E. M., De Jongh, A., Rozendaal, L., & Van Minnen, A. (2020). Trauma-focused treatment outcome for complex PTSD patients: Results of an intensive treatment programme. *European Journal of Psychotraumatology*, *11*(1), 1783955. https://doi.org/10.1080/20008198.2020.1783955

Wadji, D. L., Martin-Soelch, C., & Camos, V. (2022). Can working memory account for EMDR efficacy in PTSD? *BMC Psychology*, *10*(1), 245. https://doi.org/10.1186/s40359-022-00951-0

Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., Marrero, J., Zhu, J., & Higgins, P. D. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open*, *3*(8), e002847. https://doi.org/10.1136/bmjopen-2013-002847

Wampold, B. E. (2015). How important are the common factors in psychotherapy? An update. *World Psychiatry*, *14*(3), 270–277. https://doi.org/10.1002/wps.20238

Wampold, B. E. (2019). A smorgasbord of PTSD treatments: What does this say about integration? *Journal of Psychotherapy Integration*, *29*(1), 65–71. https://doi.org/10.1037/int0000137

Webb, C. A., Cohen, Z. D., Beard, C., Forgeard, M., Peckham, A., & Björgvinsson, T. (2020). Personalized Prognostic Prediction of Treatment Outcome for Depressed Patients in a Naturalistic Psychiatric Hospital Setting: A Comparison of Machine Learning Approaches. *Journal of Consulting and Clinical Psychology*, *88*(1), 25–38. https://doi.org/10.1037/ccp0000451

Weiss, D. S. (2007). The impact of event scale: Revised. *Cross-Cultural Assessment of Psychological Trauma and PTSD*, 219–238. https://link.springer.com/chapter/10.1007/978-0-387-70990-1_10

Wells, A., & Colbear, J. S. (2012). Treating Posttraumatic Stress Disorder With Metacognitive Therapy: A Preliminary Controlled Trial. *Journal of Clinical Psychology*, *68*(4), 373–381. https://doi.org/10.1002/jclp.20871

Wells, A., & Sembi, S. (2004). Metacognitive therapy for PTSD: A core treatment manual. *Cognitive and Behavioral Practice*, *11*(4), 365–377. https://doi.org/10.1016/S1077-7229(04)80053-1

Wells, A., Walton, D., Lovell, K., & Proctor, D. (2015). Metacognitive Therapy Versus Prolonged Exposure in Adults with Chronic Post-traumatic Stress Disorder: A Parallel Randomized Controlled Trial. *Cognitive Therapy and Research*, *39*(1), 70–80. https://doi.org/10.1007/s10608-014-9636-6

Wiggins, B. J., & Christopherson, C. D. (2019). The replication crisis in psychology: An overview for theoretical and philosophical psychology. *Journal of Theoretical and Philosophical Psychology*, *39*(4), 202-217. https://doi.org/10.1037/teo0000137

World Health Organization. (1993). *The ICD-10 Classification of Mental and Behavioural Disorders: Clinical descriptions and diagnostic guidelines*. Author. https://www.who.int/publications/i/item/9241544228

World Health Organization. (2021). *International Classification of Diseases, Eleventh Revision (ICD-11)*. https://icd.who.int/browse/2024-01/mms/en#2070699808

World Health Organization. (2024). *United Kingdom of Great Britain and Northern Ireland [Country overview]*. Data.Who.Int. https://data.who.int/countries/826

World Health Organization [WHO]. (2019). F43.1 Post-traumatic stress Disorder. In *International Statistical Classification of Diseases and Related Health Problems* (10th ed.). https://icd.who.int/browse10/2019/en#/F43

World Health Organization [WHO]. (2024). 6B41 Complex post traumatic stress disorder. In *International Classification of Diseases 11th Revision* (11th ed.). https://icd.who.int/browse/2024-01/mms/en#585833559

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Yehuda, R., Hoge, C. W., McFarlane, A. C., Vermetten, E., Lanius, R. A., Nievergelt, C. M., Hobfoll, S. E., Koenen, K. C., Neylan, T. C., & Hyman, S. E. (2015). Post-traumatic stress disorder. *Nature Reviews Disease Primers*, *1*(1), 1–22. https://doi.org/10.1038/nrdp.2015.57

Yehuda, R., & McFarlane, A. C. (1995). Conflict between current knowledge about posttraumatic stress disorder and its original conceptual basis. *Am. J. Psychiatry*, *152*, 1705–1713. https://doi.org/10.1176/ajp.152.12.1705

Yunitri, N., Chu, H., Kang, X. L., Jen, H.-J., Pien, L.-C., Tsai, H.-T., Kamil, A. R., & Chou, K.-R. (2022). Global prevalence and associated risk factors of posttraumatic stress disorder during COVID-19 pandemic: A meta-analysis. *International Journal of Nursing Studies*, *126*, 104136. https://doi.org/10.1016/j.ijnurstu.2021.104136

Zhang, H., & Zhang, Z. (1999). Feedforward networks with monotone constraints. *IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No.99CH36339)*, *3*, 1820–1823 vol.3. https://doi.org/10.1109/IJCNN.1999.832655

Zhang, J., Dashti, S. G., Carlin, J. B., Lee, K. J., & Moreno-Betancur, M. (2023). Should multiple imputation be stratified by exposure group when estimating causal effects via outcome regression in observational studies? *BMC Medical Research Methodology*, *23*(1), 42. https://doi.org/10.1186/s12874-023-01843-6

Zhang, Y., Wu, W., Toll, R. T., Naparstek, S., Maron-Katz, A., Watts, M., Gordon, J., Jeong, J., Astolfi, L., Shpigel, E., Longwell, P., Sarhadi, K., El-Said, D., Li, Y., Cooper, C., Chin-

Fatt, C., Arns, M., Goodkind, M. S., Trivedi, M. H., … Etkin, A. (2021). Identification of psychiatric disorder subtypes from functional connectivity patterns in resting-state electroencephalography. *Nature Biomedical Engineering*, *5*(4), 309–323. https://doi.org/10.1038/s41551-020-00614-8

Zhutovsky, P., Thomas, R., Olff, M., van Rooij, S., Kennis, M., van Wingen, G., & Geuze, E. (2019). Individual prediction of psychotherapy outcome in posttraumatic stress disorder using neuroimaging data. *Translational Psychiatry*, *9*(1), 326. https://doi.org/10.1038/s41398-019-0663-7

Zilcha-Mano, S., Zhu, X., Suarez-Jimenez, B., Pickover, A., Tal, S., Such, S., Marohasy, C., Chrisanthopoulos, M., Salzman, C., Lazarov, A., Neria, Y., & Rutherford, B. R. (2020). Diagnostic and Predictive Neuroimaging Biomarkers for Posttraumatic Stress Disorder. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *5*(7), 688–696. https://doi.org/10.1016/j.bpsc.2020.03.010

Zoellner, L. A., Roy-Byrne, P. P., Mavissakalian, M., & Feeny, N. C. (2019). Doubly Randomized Preference Trial of Prolonged Exposure Versus Sertraline for Treatment of PTSD. *American Journal of Psychiatry*, *176*(4), 287–296. https://doi.org/10.1176/appi.ajp.2018.17090995

Zou, H., & Hastie, T. (2005). Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

APPENDIX A

Empirically Supported Trauma-focussed Psychological Therapies

APPENDIX A - Table 1

Empirically Supported Trauma-focussed Psychological Therapies for Post-traumatic Stress Disorder

Approach	**Theory**	**Techniques**
Cognitive Processing Therapy (CPT)	Based on information processing (Foa et al., 1989) and cognitive (Beck & Emery, 1985) theories of PTSD, the theory underlying CPT is that information relating to fear is stored in cognitive networks of stimuli, responses, and meaning, which trigger avoidance behaviour. Trauma often conflicts with pre-existing schemas and this conflict is what leads to PTSD. Traumatised individuals attempt to integrate trauma with pre-existing schemas via *accommodation*, *over-accommodation*, and *assimilation*. Accommodation is where prior beliefs are adjusted to fit new information (e.g., "sometimes bad things happen despite our best efforts"), over-accommodation is where prior beliefs are significantly altered to try and prevent future traumatic experiences (e.g., "the world is a dangerous place where	The original CPT manual included three components: psychoeducation (on information processing theory), imaginal exposure (through written narrative of trauma), and cognitive reappraisal (to identify, challenge, and modify maladaptive beliefs and thought patterns). Imaginal exposure was excluded from later versions following a dismantling study which found that recounting trauma narratives did not improve outcomes and slowed therapeutic progress (Resick et al., 2008). The revised CPT protocol consists of 12 hour-long sessions, but this may vary (Resick et al., 2017).

	bad things happen frequently"), and assimilation is where new information is interpreted in a way that is congruent with prior beliefs (e.g., "what happened was my fault, I could have prevented it if I had tried harder"). The aim of CPT is to identify over-accommodated and assimilated beliefs and restructure them into accommodated beliefs through cognitive reappraisal, so that they are integrated with prior beliefs in a way that does not fundamentally change one's perception of the self or the world (Resick & Schnicke, 1992).	
Cognitive Therapy for PTSD (CT-PTSD)	Based on Ehlers and Clark's (2000) cognitive model of PTSD. In this model, high levels of arousal during trauma disrupts autobiographical memory leading to poor contextualisation and elaboration of trauma memories. As such, trauma memories are not remembered as events in the past but are re-experienced as though they are happening in the present, and this can be triggered by non-threatening cues that are in some way associated with the traumatic event. Individuals with PTSD appraise trauma and/or its	CT-PTSD begins with a personalised case formulation using Ehlers and Clark's (2000) model. Trauma memories are updated by accessing memories of the most painful parts of the trauma and their presently threatening meanings, and then updating the meanings using different information from the traumatic event or through cognitive restructuring. *Discrimination training* with triggers of reexperiencing symptoms facilitates distinction between *then* (the traumatic situation) and *now* (a safe situation). Strategies are applied to reduce

	sequelae in an excessively negative way, which produces a persistent sense of severe, present threat. This leads to behavioural and cognitive strategies to avoid perceived threat or ameliorate the sense of threat, which inadvertently maintain and exacerbate PTSD.	unhelpful behaviours and cognitions, often through behavioural experiments. *Reclaiming your life* assignments aim to regain activities and relationships lost since trauma, to address perceived permanent consequences of trauma (Ehlers et al., 2005; Ehlers & Clark, 2000). Duration is typically up to 12 weekly sessions of 60-90 minutes, followed by 3 optional monthly booster sessions (Ehlers & Wild, 2022).
Eye Movement Desensitisation and Reprocessing (EMDR)	EMDR is based on the adaptive information processing model (Shapiro, 2018). In this model, conscious and unconscious experience is processed by the brain's information processing system and integrated into interconnected memory networks of cognitive, emotional, and sensory information. This system learns from experience and guides behavioural responses to the environment. Traumatic experiences overwhelm the information processing system, preventing their adaptive processing. Memories of traumatic experiences are stored in an isolated, dysfunctional way, with their contemporaneous emotions, sensations, and perceptions intact. Current experiences can trigger re-experiencing	EMDR follows a standardised eight phase protocol. Phase 1 is case formulation and treatment planning; Phase 2 is preparation, including developing therapeutic alliance, psychoeducation, and emotion regulation skills; Phase 3 is assessment, target memories are identified, along with their associated emotions, beliefs and sensations, patients give a subjective rating of level of distress (on a 0-10 Likert scale) and identify alternative desirable beliefs; Phase 4 is desensitisation via a dual attention task (described in the adjacent column); Phase 5 is installation, once the subjective level of distress associated with the memory reaches 0, the alternative desirable belief is strengthened, which

	of some (or all) components of the trauma memory, and dysfunctionally stored memories lead to maladaptive beliefs and behaviours. The aim of EMDR is to facilitate connections between trauma memories and adaptive networks, in order to contextualise trauma and desensitise triggers (Shapiro & Maxfield, 2002). A central component of EMDR is dual focus of attention. Patients focus on an image from a traumatic memory while simultaneously engaging with an external task, typically some form of bilateral stimulation (e.g., side-to-side eye movements). This reduces the intensity of traumatic memories and their associated physiological effects and thereby aids desensitisation and reprocessing. Several theories have been proposed to explain the mechanism(s) through which bilateral stimulation aids this process, with varying degrees of empirical support. These include processes related to rapid eye movement sleep, episodic memory and interaction between the left and right hemispheres of the brain, and, most promisingly, working memory capacity	contradicts negative beliefs; Phase 6 is body scan, in which patients focus on the physical sensations within their body from head to toe, observing and releasing any residual sensations associated with the target memory; Phase 7 is closure, this takes place at the end of each session, progress is reviewed and emotional stability is re-established if necessary; Phase 8 is re-evaluation, the effectiveness of the treatment episode is assessed and future treatment is planned. EMDR sessions are typically 50-90 minutes, each treatment phase can span multiple sessions, and phases can be repeated for different traumatic memories (Shapiro, 2018).

	and the amygdala (de Jongh et al., 2024; Landin-Romero et al., 2018; Wadji et al., 2022).	
Prolonged Exposure (PE)	Based on emotional processing theory (Foa & Kozak, 1986), which posits that fear is represented in memory as a cognitive structure of stimuli, responses, and associated meaning. Trauma is not emotionally processed during the event, producing dysfunctional fear structures that trigger physiological and escape/avoidance fear responses to non-threatening stimuli (Foa & Cahill, 2001). Prolonged exposure aims to amend dysfunctional fear structures, by first activating the fear structure through in vivo and/or imaginal exposure and then introducing new information that is incongruent with the dysfunctional associations in the fear structure.	PE consists of three key components: psychoeducation (to establish the rationale for PE), in vivo exposure (approaching safe situations, places, or people associated with trauma), and imaginal exposure (approaching thoughts, emotions, memories associated with trauma) followed by processing. To facilitate imaginal exposure, patients recount traumatic experiences out loud and in the present tense. This is followed by processing, in which the patient and therapist discuss and interpret the emotions and perceptions that arose during imaginal exposure. Duration is typically 8-15 once-or-twice-weekly 90-minute sessions (McLean & Foa, 2024).

Note. PTSD = post-traumatic stress disorder.

APPENDIX B
Database Search Terms

Scopus

((TITLE-ABS-KEY("psychotherapy" OR "psychological therapy" OR "Cognitive behavio* therapy" OR "cognitive therapy" OR "CBT" OR "cognitive processing therapy" OR "Exposure therapy" OR "Prolonged exposure" OR "narrative exposure" OR "Eye Movement Desensiti*ation and Reprocessing" OR "EMDR" OR "trauma focussed therapy" OR "Brief Eclectic Psychotherapy"))

AND

(TITLE-ABS-KEY("posttraumatic" OR "post traumatic" OR "post-traumatic" OR "traumatic stress" OR "traumatic memor*" OR "ptsd" OR "cptsd"))

AND

(TITLE-ABS-KEY("machine learning" OR "machine-learning" OR "supervised learning" OR "unsupervised learning" OR algorithm OR "statistical learning" OR "artificial intelligence" OR "AI" OR "data mining" OR "deep learning" OR "kernel" OR "personali*ed advantage ind*" OR "regularized" OR "ridge" OR "least absolute shrinkage and selection operator" OR "LASSO" OR "elastic net" OR "decision tree*" OR "random forest*" OR "regression tree*" OR "classification tree*" OR "nearest neighb*" OR "k-nn" OR "k-means" OR "support vector" OR "vector machine" OR "SVM" OR "SVR" OR "naïve Bayes" OR "Bayesian network" OR "neural network" OR "perceptron" OR "radial basis function" OR "cluster analysis" OR "principal components analysis" OR "latent transition" OR "autoencoder" OR "dimensionality reduction" OR "latent Dirichlet allocation" OR "LDA"

OR "chi-square automatic interaction detection" OR "CHAID" OR "XGBoost" OR "bootstrap*" OR "bagging" OR "boosting")))

PubMed

(("psychotherapy"[MeSH Terms] OR "psychotherapy"[Title/Abstract] OR "Cognitive behavio* therapy"[Title/Abstract] OR "cognitive therapy"[Title/Abstract] OR "CBT"[Title/Abstract] OR "cognitive processing therapy"[Title/Abstract] OR "Exposure therapy"[Title/Abstract] OR "Prolonged exposure"[Title/Abstract] OR "narrative exposure"[Title/Abstract] OR "Eye Movement Desensiti*ation and Reprocessing"[Title/Abstract] OR "EMDR"[Title/Abstract] OR "trauma focused therapy"[Title/Abstract] OR "Brief Eclectic Psychotherapy"[Title/Abstract])

AND

("Stress Disorders, Post-Traumatic"[MeSH Terms] OR "post-traumatic"[Title/Abstract] "posttraumatic"[Title/Abstract] OR "post traumatic"[Title/Abstract] OR "traumatic stress"[Title/Abstract] OR "traumatic memor*"[Title/Abstract] OR "ptsd"[Title/Abstract] OR "cptsd"[Title/Abstract])

AND

(("machine learning"[MeSH Terms] OR "machine-learning"[Title/Abstract] OR "machine learning"[Title/Abstract] OR "supervised learning"[Title/Abstract] OR "unsupervised learning"[Title/Abstract] OR "algorithm"[Title/Abstract] OR "statistical learning"[Title/Abstract] OR "artificial intelligence"[Title/Abstract] OR "AI"[Title/Abstract] OR "data mining"[Title/Abstract] OR "deep learning"[Title/Abstract] OR "kernel"[Title/Abstract] OR "personali*ed advantage ind*"[Title/Abstract] OR "regularized"[Title/Abstract] OR "ridge"[Title/Abstract] OR "least absolute shrinkage and selection operator"[Title/Abstract] OR "LASSO"[Title/Abstract] OR "elastic net"[Title/Abstract] OR "decision tree*"[Title/Abstract] OR "random forest*"[Title/Abstract] OR "regression tree*"[Title/Abstract] OR "classification tree*"[Title/Abstract] OR "nearest neighb*"[Title/Abstract] OR "k-nn"[Title/Abstract] OR "k-means"[Title/Abstract] OR "support vector"[Title/Abstract] OR "vector machine"[Title/Abstract] OR "SVM"[Title/Abstract] OR "SVR"[Title/Abstract] OR "naïve Bayes"[Title/Abstract] OR "Bayesian network"[Title/Abstract] OR "neural network"[Title/Abstract] OR "perceptron"[Title/Abstract] OR "radial basis function"[Title/Abstract] OR "cluster analysis"[Title/Abstract] OR "principal components analysis"[Title/Abstract] OR "latent transition"[Title/Abstract] OR "autoencoder"[Title/Abstract] OR "dimensionality reduction"[Title/Abstract] OR "latent Dirichlet allocation"[Title/Abstract] OR "LDA"[Title/Abstract] OR "chi-square automatic interaction detection"[Title/Abstract] OR "CHAID"[Title/Abstract] OR "XGBoost"[Title/Abstract] OR "bootstrap*"[Title/Abstract] OR "bagging"[Title/Abstract] OR "boosting"[Title/Abstract])))

APA PsycInfo via Ovid

(("psychotherapy" OR "psychological therapy" OR "Cognitive behavio$ therapy" OR "cognitive therapy" OR "CBT" OR "Exposure therapy" OR "Prolonged exposure" OR "narrative exposure" OR "Eye Movement Desensiti$ation and Reprocessing" OR "EMDR" OR "trauma focussed therapy" OR "Brief Eclectic Psychotherapy").ab,ti,id.

AND

("posttraumatic" OR "post traumatic" OR "traumatic stress" OR "traumatic memor$" OR "ptsd" OR "cptsd").ab,ti,id.

AND

("machine learning" OR "machine-learning" OR "supervised learning" OR "unsupervised learning" OR algorithm OR "statistical learning" OR "artificial intelligence" OR "AI" OR "data mining" OR "deep learning" OR "kernel" OR "personali*ed advantage ind*" OR "regularized" OR "ridge" OR "least absolute shrinkage and selection operator" OR "LASSO" OR "elastic net" OR "decision tree*" OR "random forest*" OR "regression tree*" OR "classification tree*" OR "nearest neighb*" OR "k-nn" OR "k-means" OR "support vector" OR "vector machine" OR "SVM" OR "SVR" OR "naïve Bayes" OR "Bayesian network" OR "neural network" OR "perceptron" OR "radial basis function" OR "cluster analysis" OR "principal components analysis" OR "latent transition" OR "autoencoder" OR "dimensionality reduction" OR "latent Dirichlet allocation" OR "LDA" OR "chi-square automatic interaction detection" OR "CHAID" OR "XGBoost" OR "bootstrap*" OR "bagging" OR "boosting").ab,ti,id.)

PTSDpubs via ProQuest

(NOFT(("psychotherapy" OR "psychological therapy" OR "Cognitive behavio* therapy" OR "CBT" OR "cognitive therapy" OR "cognitive processing therapy" OR "Exposure therapy" OR "Prolonged exposure" OR "narrative exposure" OR "Eye Movement Desensiti*ation and Reprocessing" OR "EMDR" OR "trauma focussed therapy" OR "Brief Eclectic Psychotherapy") AND ("posttraumatic" OR "post traumatic" OR "post-traumatic" OR "traumatic stress" OR "traumatic memor*" OR "ptsd" OR "cptsd") AND ("machine learning" OR "machine-learning" OR "supervised learning" OR "unsupervised learning" OR algorithm OR "statistical learning" OR "artificial intelligence" OR "AI" OR "data mining" OR "deep learning" OR "kernel" OR "personali*ed advantage ind*" OR "regularized" OR "ridge" OR "least absolute shrinkage and selection operator" OR "LASSO" OR "elastic net" OR "decision tree*" OR "random forest*" OR "regression tree*" OR "classification tree*" OR "nearest neighb*" OR "k-nn" OR "k-means" OR "support vector" OR "vector machine" OR "SVM" OR "SVR" OR "naïve Bayes" OR "Bayesian network" OR "neural network" OR "perceptron" OR "radial basis function" OR "cluster analysis" OR "principal components analysis" OR "latent transition" OR "autoencoder" OR "dimensionality reduction" OR "latent Dirichlet allocation" OR "LDA" OR "chi-square automatic interaction detection" OR "CHAID" OR "XGBoost" OR "bootstrap*" OR "bagging" OR "boosting")))

APPENDIX C

Studies Excluded During Full Text Screening

APPENDIX C - Table 1

Studies Excluded During Full Text Screening

Author	Title	DOI	Reason For Exclusion
Barnes et al. (2019)	Developing predictive models to enhance clinician prediction of suicide attempts among veterans with and without PTSD	https://doi.org/10.1111/sltb.12511	Did not predict PTSD treatment outcome (predicted suicide)
Bryant et al. (2008)	Amygdala and ventral anterior cingulate activation predicts treatment response to cognitive behaviour therapy for post-traumatic stress disorder	https://doi.org/10.1017/s0033291707002231	No machine learning methods
de Kleine et al. (2014)	Prescriptive variables for d-cycloserine augmentation of exposure therapy for posttraumatic stress disorder	https://doi.org/10.1016/j.jpsychires.2013.10.008	No machine learning methods
Dorrepaal et al. (2013)	Treatment compliance and effectiveness in complex PTSD patients with co-morbid personality disorder undergoing stabilizing cognitive behavioral group treatment: A preliminary study	https://doi.org/10.3402/ejpt.v4i0.21171	Psychological therapy for PTSD not recommended by CPG (group CBT only)
Galatzer-Levy et al. (2014)	Quantitative forecasting of PTSD from early trauma responses: A machine learning application	https://doi.org/10.1016/j.jpsychires.2014.08.017	Did not predict treatment outcome (predicted onset of PTSD)
Galatzer-Levy et al. (2017)	Utilization of machine learning for prediction of post-traumatic stress: A re-examination of cortisol in the prediction and pathways to non-remitting PTSD	https://doi.org/10.1038/tp.2017.38	Did not predict treatment outcome (predicted onset of PTSD)

Hilbert et al. (2021)	Identifying CBT non-response among OCD outpatients: A machine-learning approach	https://doi.org/10.1080/10503307.2020.1839140	Treatment not for PTSD (treatment for OCD)
Karstoft et al. (2015)	Bridging a translational gap: Using machine learning to improve the prediction of PTSD	https://doi.org/10.1186/s12888-015-0399-8	Did not predict treatment outcome (predicted onset of PTSD)
Korgaonkar et al. (2020)	Intrinsic connectomes underlying response to trauma-focused psychotherapy in post-traumatic stress disorder	https://doi.org/10.1038/s41398-020-00938-8	No machine learning methods
Lutz et al. (2019)	Towards integrating personalized feedback research into clinical practice: Development of the Trier Treatment Navigator (TTN)	https://doi.org/10.1016/j.brat.2019.103438	Sample not adults with PTSD (only 5.7% of participants were seeking treatment for PTSD)
Rizvi et al. (2009)	Cognitive and affective predictors of treatment outcome in Cognitive Processing Therapy and Prolonged Exposure for posttraumatic stress disorder	https://doi.org/10.1016/j.brat.2009.06.003	No machine learning methods
Roberge et al. (2019)	Predicting response to cognitive processing therapy: Does trauma history matter?	https://doi.org/10.1037/tra0000530	No machine learning methods
Rubel et al. (2019)	Predicting personalized process-outcome associations in psychotherapy using machine learning approaches—A demonstration	https://doi.org/10.1080/10503307.2019.1597994	Sample not adults with PTSD (only 3.9% of the sample sought treatment for PTSD)
Schultebraucks et al. (2021)	Pre-deployment risk factors for PTSD in active-duty personnel deployed to Afghanistan: A machine-learning approach for analyzing multivariate predictors	https://doi.org/10.1038/s41380-020-0789-2	Did not predict treatment outcome (predicted onset of PTSD)
Schultebraucks et al. (2020)	A validated predictive algorithm of post-traumatic stress course following emergency department admission after a traumatic stressor	https://doi.org/10.1038/s41591-020-0951-z	Did not predict treatment outcome (predicted onset of PTSD)

Schumm et al. (2013)	Latent class differences explain variability in PTSD symptom changes during cognitive processing therapy for veterans	https://doi.org/10.1037/a0030359	No machine learning methods
Senger et al. (2021)	Predicting optimal treatment outcomes using the Personalized Advantage Index for patients with persistent somatic symptoms	https://doi.org/10.1080/10503307.2021.1916120	Treatment not for PTSD (treatment for persistent somatic symptoms)
Smith et al. (2019)	Predictors of dropout from residential treatment for posttraumatic stress disorder among military veterans	https://doi.org/10.3389/fpsyg.2019.00362	No machine learning methods
Sonne et al. (2016)	Psychosocial predictors of treatment outcome for trauma-affected refugees	https://doi.org/10.3402/ejpt.v7.30907	No machine learning methods
Stein et al. (2012)	Trajectories of response to treatment for posttraumatic stress disorder	https://doi.org/10.1016/j.beth.2012.04.003	No machine learning methods
Taylor (2003)	Outcome predictors for three PTSD treatments: Exposure therapy, EMDR, and relaxation training	https://dx.doi.org/10.1891/jcop.17.2.149.57432	No machine learning methods
Taylor et al. (2001)	Posttraumatic stress disorder arising after road traffic collisions: Patterns of response to cognitive-behavior therapy	https://doi.org/10.1037/0022-006X.69.3.541	Psychological therapy for PTSD not recommended by CPG (group CBT)
van Geusau et al. (2021)	Predicting outcome in an intensive outpatient PTSD treatment program using daily measures	https://doi.org/10.3390/jcm10184152	No machine learning methods
van Minnen et al. (2002)	Prolonged exposure in patients with chronic PTSD: Predictors of treatment outcome and dropout	https://doi.org/10.1016/s0005-7967(01)00024-9	No machine learning methods
van Rooij et al. (2014)	Neural correlates of inhibition and contextual cue processing related to treatment response in PTSD	https://doi.org/10.1038/npp.2014.220	No machine learning methods

van Rooij et al. (2015)	Predicting treatment outcome in PTSD: A longitudinal functional MRI study on trauma-unrelated emotional processing	https://doi.org/10.1038/npp.2015.257	No machine learning methods
Vöhringer et al. (2020)	Should I stay or must I go? Predictors of dropout in an internet-based psychotherapy programme for posttraumatic stress disorder in Arabic	https://doi.org/10.1080/20008198.2019.1706297	Psychological therapy for PTSD not recommended by CPG (Internet-based CBT)
Wester et al. (2022)	Covariate selection for estimating individual treatment effects in psychotherapy research: A simulation study and empirical example	https://doi.org/10.1177/21677026211071043	Sample not adults with PTSD (Simulation study)
Yuan et al. (2018)	Pre-treatment resting-state functional MR imaging predicts the long-term clinical outcome after short-term Paroxtine treatment in post-traumatic stress disorder	https://doi.org/10.3389/fpsyt.2018.00532	No psychological therapy (Pharmacological treatment only)
Zegerius et al. (2021)	Modelling metaplasticity and memory reconsolidation during an eye-movement desensitization and reprocessing treatment	https://doi.org/10.1007/978-3-030-65596-9_74	No machine learning methods
Zhutovsky et al. (2021)	Individual prediction of trauma-focused psychotherapy response in youth with posttraumatic stress disorder using resting-state functional connectivity	https://doi.org/10.1016/j.nicl.2021.102898	Sample not adults with PTSD (Sample < 18 years old)

Note. CBT = cognitive behavioural therapy; CPG = clinical practice guidelines; OCD = obsessive compulsive disorder; PTSD = post-traumatic stress disorder.

APPENDIX D

Risk of Bias Assessments

APPENDIX D - Table 1

Risk of Bias Assessments

	PROBAST Signalling Question	Study																
		Deisenhofer et al. (2018)	Etkin et al. (2019)	Fleming et al. (2018)	Forbes et al. (2003)	Held et al. (2022)	Hendriks et al. (2018)	Herzog et al. (2021)	Hoeboer et al. (2021)	Keefe et al. (2018)	Kratzer et al. (2019)	López-Castro et al. (2021)	Nixon et al. (2021)	Stirman et al. (2021)	Stuke et al. (2021)	Zhang et al. (2021)	Zhutovsky et al. (2019)	Zilcha-Mano et al. (2020)
1.1	Were appropriate data sources used, e.g., cohort, RCT or nested case-control study data?	N	Y	N	N	Y	Y	N	Y	Y	N	Y	Y	Y	N	Y	Y	Y
1.2	Were all inclusions and exclusions of participants appropriate?	PY	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Risk of bias introduced by selection of participants	High	Low	High	High	Low	Low	High	Low	Low	High	Low	Low	Low	High	Low	Low	Low
2.1	Were predictors defined and assessed in a similar way for all participants?	PY	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
2.2	Were predictor assessments made without knowledge of outcome data?	Y	PY	PY	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
2.3	Are all predictors available at the time the model is intended to be used?	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y

	Risk of bias introduced by predictors or their assessment	Low	Low	Low	Low	Low	High	Low	Low	Low	Low	High	Low	Low	Low	Low	Low	Low
3.1	Was the outcome determined appropriately?	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
3.2	Was a pre-specified or standard outcome definition used?	Y	PY	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
3.3	Were predictors excluded from the outcome definition?	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
3.4	Was the outcome defined and determined in a similar way for all participants?	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
3.5	Was the outcome determined without knowledge of predictor information?	Y	NI	Y	Y	Y	Y	Y	PY	Y	Y	Y	Y	Y	Y	Y	Y	Y
3.6	Was the time interval between predictor assessment and outcome determination appropriate?	PY	PY	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Risk of bias introduced by the outcome or its determination	Low	Unclear	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low
4.1	Were there a reasonable number of participants with the outcome?	N	N	N	N	N	N	PY	N	N	N	N	N	N	N	PN	N	N
4.2	Were continuous and categorical predictors handled appropriately?	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
4.3	Were all enrolled participants included in the analysis?	N	N	N	N	N	N	Y	Y	N	N	N	N	N	N	Y	N	Y
4.4	Were participants with missing data handled appropriately?	Y	NI	NI	PY	N	Y	Y	Y	Y	Y	NI	N	Y	N	NI	NI	NI

4.5	Was selection of predictors based on univariable analysis avoided?	Y	N	Y	Y	Y	Y	Y	Y	N	N	Y	Y	Y	Y	Y	N	Y
4.6	Were complexities in the data (e.g., censoring, competing risks, sampling of controls) accounted for appropriately?	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
4.6	Were relevant model performance measures evaluated appropriately?	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	NA
4.8	Were model overfitting and optimism in model performance accounted for?	Y	Y	N	N	Y	N	Y	Y	Y	N	N	Y	Y	Y	N	Y	Y
4.9	Do predictors and their assigned weights in the final model correspond to the results from multivariable analysis?	Y	NI	Y	NI	NI	Y	Y	Y	Y	Y	N	NI	Y	NI	NI	NI	NA
	Risk of bias introduced by the analysis	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High
	Overall risk of bias assessment	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High

Note. N = no; NA = not applicable; NI = no information; PN = probably no; PY = probably yes; Y = yes.

APPENDIX E

Findings of Studies Included in Systematic Review

APPENDIX E - Table 1

Study Results and Study Author's Interpretation

Study	Predictors in Final Model(s) (*n* Predictors in Model)	Evaluation Metrics Reported	Findings	Study Author's Interpretation

Deisenhofer et al. (2018)	*Tf-CBT:*
	Functional impairment (WSAS), employment status, age, gender
	(4)
	
	EMDR:
	Depression (PHQ-9), medication
	(2)
		True error:
		Tf-CBT = 4.92
		EMDR = 5.37
		Whole analysed sample = 5.07
			For Tf-CBT, higher functional impairment (WSAS; $\beta = 0.32$, $p < .001$), being male ($\beta = -0.14$, $p \geq .05$), being unemployed ($\beta = -0.33$, $p < .001$), and younger age ($\beta = -0.17$, $p < .05$) predicted more severe symptoms post-treatment.
			For EMDR, higher baseline depression (PHQ-9; $\beta = 0.40$, $p < .01$) and being prescribed antidepressants ($\beta = 0.29$, $p < .001$) predicted more severe symptoms post-treatment.
			Patients who received their model indicated optimal treatment ($n = 124$) reported significantly better outcomes than those who received their model indicated suboptimal treatment ($n = 101$; Cohen's $d = 0.40$, 95% CI [0.13, 0.67], number needed to treat = 4.49).
			There was a significantly higher rate of reliable improvement among patients who received their model indicated optimal treatment (62.9%) versus those who received their suboptimal treatment (33.66%; $\chi^2(1, n = 225) = 19.54$, $p < .001$).
				Results consistent with previous literature.
				Limitations: Exploratory modelling using retrospective data, small sample, and propensity score matching, no PTSD symptom measure.
				Prospective testing is required with a larger sample and PTSD symptom outcome measure.
				Integrating a PAI into routine clinical practice could potentially improve PTSD outcomes by guiding selection of optimal treatment.

Etkin et al. (2019)	Verbal memory delayed recall impairment,
	within ventral attention network functional connectivity
	(2)
		Linear support vector machine:
		Accuracy = 85%
		Sensitivity = 80%
		Specificity = 87%
		
		Radial basis function support vector machine:
		Accuracy = 90%
		Sensitivity = 80%
		Specificity = 93%
			Generalised linear model results indicated that impaired verbal memory, low within Ventral Attention Network connectivity, and their interaction, were not significantly associated with baseline PTSD severity, depression severity, PTSD symptom clusters, dissociative symptoms, comorbid diagnoses, alcohol use, traumatic brain injury, or quality of life.
			Generalised linear mixed model results indicated that the interaction between verbal memory recall and within Ventral Attention Network functional connectivity moderated treatment outcome (PE vs. wait list), in that impaired verbal memory and low within Ventral Attention Network connectivity predicted no change in PTSD symptom score in the PE group.
				Findings consistent with recent calls to move away from symptom focussed definitions of PTSD heterogeneity to cognitive-neurological definitions.
				Replication is needed.

Study	Predictors in Final Model(s) (*n* Predictors in Model)	Evaluation Metrics Reported	Findings	Study Author's Interpretation
			Support vector machine results indicated that the combination of low within Ventral Attention Network connectivity and impaired verbal memory predicted non-response to PE. Either variable alone did not predict outcome (accuracies ≤ 65%, *p* > 0.18).	
Fleming et al. (2018)	Days between introductory session and therapy invitation, URICA item 9 (“*I have been successful in working on my trauma issues, but I’m not sure I can keep up the effort on my own*”), URICA item 5 (“*I’m not the one with trauma issues, it doesn’t make much sense for me to be here*”), presence of traumatic brain injury, prior PTSD treatment (5)	None	Wait time before treatment was the best predictor of attendance, followed by motivational readiness to address trauma, then traumatic brain injury and prior PTSD treatment. Exhaustive CHAID classification tree indicated that participants who waited > 68 days to begin treatment attended significantly fewer treatment sessions; and among those who waited > 68 days, participants who disagreed with the URICA scale item 9 attended significantly fewer sessions. Among those who waited ≤ 68 days, participants who strongly disagreed with URICA item 5 completed significantly more sessions; among those who strongly disagreed with URICA item 5, those with traumatic brain injury attended significantly more sessions; and among those who didn't strongly disagree with URICA item 5, participants with no prior PTSD treatment attended significantly more sessions.	The finding that shorter wait time and motivation is associated with greater retention is congruent with previous research. However, contrary to previous research and theory, the current findings suggest that these are most important predictors of engagement above demographic and clinical variables; this requires further study and replication. This study was exploratory, and the use of naturalistic clinical data limits internal validity. There may be extraneous variables associated with treatment retention. Clinical implications: Engage veterans quickly and focus on motivation and preparation for treatment.
Forbes et al. (2003)	Clinical subscales and validity subscales of MMPI-2, anxiety and depression (HADS), alcohol use (AUDIT) (16)	None	Ward's cluster analysis identified three groups, and this was supported by the *k*-means cluster analysis with three-cluster classification agreement in 125 of the 158 cases (*n* = 158, alpha = 0.68, *p* < 0.001). Group 1 (*n* = 62) scored highly on PTSD symptom severity, but lower in other psychiatric measures; Group 2 (*n* = 38) scored significantly lower on PTSD symptom severity, and low on other psychiatric measures; and Group 3 (*n* = 36) scored as highly as Group 1 on PTSD symptom severity, and significantly higher on personality disturbance, other forms of psychopathology, emotional vulnerability, and expressed distress. Repeated-measures multivariate general linear model found significant effect of time (*F*(2,132) = 22.56, *p* < 0.001) and group (*F*(2,133) = 12.51, *p* < 0.001), and a significant time by group interaction (*F*(4,266) = 3.35, *p* < 0.02). Univariate post-hoc analyses revealed that Group 1 and 3 significantly improved from baseline to 3 months, but Group 2 did not. Cross-sectional analyses revealed a	Based on MMPI-2 characteristics groups may be labelled as follows: Group 1 = High PTSD-introversion/ somatization. Group 2 = Low PTSD-subclinical personality pathology. Group 3 = High PTSD-disinhibition/externalization. Group 2 represents a "subclinical" group, with lower intake PTSD, anxiety, and depression scores than Groups 1 and 3. Groups 1 and 3 reported similarly severe intake symptom scores, but Group 3 scored higher on Factor 2 (indicating greater personality disturbance and psychopathology), and lower on Factor 3 (indicating greater emotional vulnerability and expressed distress).

Study	Predictors in Final Model(s) (*n* Predictors in Model)	Evaluation Metrics Reported	Findings	Study Author's Interpretation
			significant difference between groups mean PCL scores at baseline ($F(2,133) = 15.91$, $p < 0.001$), 3 months post-treatment ($F(2,133) = 5.28$, $p < 0.01$), and 9 months post-treatment ($F(2,133) = 8.48$, $p < 0.001$); and Scheffe post-hoc analyses revealed that Group 2 had significantly less severe PTSD symptoms at baseline than Groups 1 and 3 (with no difference between 1 and 3), but at 9 month follow up Group 3 had significantly more severe PTSD symptoms than Groups 1 and 2 (with no difference between 1 and 2).	There are some similarities to/consistencies with subgroups identified in previous studies. Personality pathology and psychopathology of Group 3 may have negatively affected long term symptom change, consistent with previous research on other disorders with comorbid personality disorder. Poor treatment response of Group 2 may be due to a floor effect, mild symptoms at intake allows less capacity for change. Limitations: Imprecision of procedure of identifying optimal solution in cluster analysis; small sample size limits generalizability; no non-treatment control group, no comparison to natural pattern of symptom change over time.
Held et al. (2022)	*Gradient Boosted Models predicting Fast Response class:* Post-traumatic stress symptoms (item level responses to PCL-5 and CAPS-5), post-traumatic cognitions (item level responses to PTCI) (10) *Elastic Net predicting Minimal Response class:* Post-traumatic stress symptoms (item level responses to PCL-5 and CAPS-5), post-traumatic cognitions (item level responses to PTCI), marriage/domestic partnership status, age, gender, level of education, military service branch (60)	Predicting fast response AUC-PR: Gradient Boosted Models = 0.466 Random Forest = 0.457 MMPC LR = 0.450 Elastic Net = 0.405 Ridge Classification = 0.394 Logistic Regression = 0.224 AUC-ROC: Gradient Boosted Models = 0.765 Predicting minimal response AUC-PR: Gradient Boosted Models = 0.583 Random Forest = 0.595 MMPC LR = 0.579 Elastic Net = 0.628 Ridge Classification = 0.611	Group Based Trajectory Modelling classified *n* = 61 participants (14.1%) as fast responders, and *n* = 73 participants (16.9%) as minimal responders. When predicting fast response class membership Gradient Boosted Models performed best and selected 10 predictor variables, all of which were trauma related and were item level scores from PCL-5, PTCI, and CAPS-5 measurements. When predicting minimal response class membership Elastic Net performed best and selected 60 predictor variables. This included 7 demographic variables, but the trauma related clinical variables were the most important predictors.	These findings are congruent with those of similar, recent studies. Strengths include large sample size, large number of variables, generalizability of naturalistic clinical sample. Limitations include reliance on self-report measures, outcome class imbalance, trajectories based on probabilistic trajectory modelling (which can be sample dependent and could reduce generalizability), and the likelihood of misclassification limits clinical usefulness. It is not recommended that clinicians rely on this model alone when deciding whether to include/exclude a patient from treatment.

Study	Predictors in Final Model(s) (*n* Predictors in Model)	Evaluation Metrics Reported	Findings	Study Author's Interpretation
		Logistic Regression = 0.288 AUC-ROC: Elastic Net = 0.826		
Hendriks et al. (2018)	Living condition, between-session fear habituation (2)	None	*k*-means clustering analysis identified four response trajectories: fast responders (*n* = 9), slow responders (*n* = 18), partial responders (*n* = 22), and non-responders (*n* = 20). Living Condition was the only significant pre-treatment predictor of response trajectory class, in that living alone significantly predicted partial response cluster membership (B = -1.89, SE = .70, *p* < .05, OR = .15, 95% CI [.04, .59]). Additionally, greater between-session fear habituation significantly predicted Fast Response cluster membership (B = .76, SE = .39, *p* = .05, OR = 2.13, 95% CI [1.00, 4.54]).	The response trajectories identified corresponded to those identified in previous research. The finding that between-session, but not within-session, fear habituation was associated with outcome is coherent with previous research. The finding that patients in the partial-responders cluster were more likely to be living alone than those in the non-responders cluster conflicts with previous findings; this may be due to complex PTSD and partners facilitating PTSD-related avoidance behaviour. Early treatment process variables more robust predictors of intensive PTSD treatment outcomes than baseline demographic and clinical variables. Replication in a randomised controlled design is required. Other limitations include the low number of participants per cluster; standardised, validated measures were not used to measure treatment resistance, suicidal ideation, self-harm, aggressive behaviour, and sense of losing control; therapist adherence was measured via self-report; DSM-IV-TR (CAPS) diagnostic criteria was used instead of DSM-5 (CAPS-5).

Herzog et al. (2021)	PTSD severity (IES-R total), psychoticism (BSI), avoidance (IES-R subscale), wish to retire, depression (BDI-II), number of comorbid diagnoses, age, bronchial asthma, physical symptoms (PHQ-15), outpatient psychiatric care, children, being retired, work disability in past year, outpatient psychotherapy, somatization (BSI)
	(15)
		Training set:
		R^2 = 0.17
		MAE = 0.69
		RMSE = 0.91
		
		Test set:
		R^2 = 0.16
		MAE = 0.77
		RMSE = 0.95
			Elastic Net Regularization selected 11 clinical/psychological factors, and 4 sociodemographic factors. The strongest and most stable predictors were clinical/psychological, and the top three were baseline PTSD score (IES-R, β = .207), psychoticism (BSI subscale, β = -.110), and avoidance (IES-R subscale, β = .088). The strongest and most stable sociodemographic predictors were 'wish to retire' (β = -.078) and older age (β = -.038).
			More severe baseline PTSD (IES-R; β = .207) and PTSD-related avoidance (IES-R subscale; β = .088) were associated with better outcomes; whereas higher psychoticism (BSI subscale; β = -.110), wish to retire (β = -.078), higher baseline depression (BDI-II; β =
				A prediction model combining predictors from multiple domains (sociodemographic, clinical, psychometric) could help to predict variance in treatment outcome between different patients.
				Findings regarding predictors associated with outcome were in line with previous theory and research. The proportion of variance (R^2) explained by the model is satisfactory, but some variables that were not measured in routine care may be important predictors, such as expectations, childhood maltreatment, treatment resistance and chronicity.

Study	Predictors in Final Model(s) (*n* Predictors in Model)	Evaluation Metrics Reported	Findings	Study Author's Interpretation
			-.068), greater number of comorbid diagnoses (β = -.057), older age (β = -.038), bronchial asthma (β = -.024), being retired (β = -.007), and work disability in last year (β = -.005) were associated with poorer treatment outcomes.	The proportion of variance explained in the test set is comparable to that in the training set, suggesting that overfitting of the model to the data did not occur. Limitations: Risk of lower internal validity, exploratory, naturalistic study using retrospective data, lack of strict manual for treatment limits generalizability, no assessment of therapist effects, no follow up measures, no control group, no randomization, no blinding of outcome assessors, large percentage of patients were excluded due to missing outcome measure data, generalizability limited to inpatient multi-modal treatment setting (common in Germany but not elsewhere), did not control for concurrent psychopharmacological interventions.
Hoeboer et al. (2021)	*CAPS-5 change following PE/IPE:* Depression (BDI), social support (MOS), concurrent mental health problems (MINI axis 1), childhood sexual abuse (CTQ) (4) *CAPS-5 change following STAIR+PE:* General health status (EQ-5D-5L), emotion regulation difficulties (DERS), PTSD severity (CAPS-5) (3) *PCL-5 change following PE or IPE:* Depression (BDI), social support (MOS) (2) *PCL-5 change following STAIR and PE:* General health status (EQ-5D-5L),	RMSE (referred to as *average error* in the publication) *CAPS-5:* PE/IPE = 5.09 (SD = 7.57) STAIR+PE = 4.06 (SD = 7.25) *PCL-5:* PE/IPE = 7.09 (SD = 6.16) STAIR+PE = 7.24 (SD = 4.74)	When predicting change in observer rated PTSD symptom (CAPS-5) score following PE or IPE, higher depression score (BDI), higher childhood sexual abuse score (CTQ), lower social support score (MOS), and more DSM-IV axis-1 diagnoses (MINI), predicted poorer response to treatment. When predicting change in observer rated PTSD symptom (CAPS-5) score following STAIR+PE, higher difficulties in emotion regulation (DERS), higher baseline PTSD (CAPS-5), and lower general health status (EQ-5D-5L), predicted poorer response to treatment. With CAPS-5 as outcome, 50% of patients were randomised to their optimal treatment, patients randomised to their optimal treatment reported a significantly larger reduction in symptoms (Mean (SD) reduction = 22.96 (6.99)) than patients randomised to their suboptimal treatment (Mean (SD) reduction = 18.94 (7.57); *F*(1,147) = 11.36, *p* < 0.001; Cohen's *d* = 0.55, 95% CI [0.23, 0.88]). When predicting change in self-rated PTSD symptom (PCL-5) score following PE or IPE, higher depression (BDI) score, and lower social support (MOS) score predicted poorer response to treatment. When predicting change in self-rated PTSD symptom (PCL-5) score following STAIR+PE, lower general health status (EQ-5D-5L) score and higher difficulties in emotion regulation (DERS) scores predicted poorer response to treatment.	Approximately half of the patients were randomised to their suboptimal treatment and may have benefitted from model-based treatment selection. Predictors selected were consistent with previous personalised PTSD treatment studies, except that no demographic variables were selected as predictors. This may be due to the greater number of candidate clinical predictors, which appear to be better predictors of outcome than demographics. The finding that more emotion regulation difficulties predicted poorer outcomes in the STAIR+PE condition is in contrast to previous findings, and is important as STAIR+PE was developed for patients with severe emotion regulation difficulties who may find it difficult to tolerate PE. Many variables often identified as predictors of PTSD treatment outcomes were not selected as predictors in either model, such as dissociation and personality disorders. Limitations: Sample size did not allow for *k*-fold cross validation or partitioning of a holdout sample; PAI based on linear combination of predictors, and some predictors

Study	Predictors in Final Model(s) (*n* Predictors in Model)	Evaluation Metrics Reported	Findings	Study Author's Interpretation
	emotion regulation difficulties (DERS) (2)		With PCL-5 as outcome, 63% of patients were randomised to their optimal treatment, patients randomised to their optimal treatment reported a significantly larger reduction in symptoms (Mean (SD) reduction = 25.65 (10.4)) than patients randomised to their suboptimal treatment (Mean (SD) reduction = 20.96 (9.84); *F*(1,147) = 7.67, *p* = 0.006; Cohen's *d* = 0.47, 95% CI [0.13, 0.81]).	identified by the Boruta algorithm but dropped during bootstrapping may be non-linearly related to outcome; relationship between baseline CAPS-5 and change in CAPS-5 (outcome) may be due to regression to the mean and mathematical coupling.
Keefe et al. (2018)	*Prescriptive variables:* Childhood physical abuse, current relationship abuse, trait anger, race *Prognostic variables:* Years of education, estimated IQ score (6)	None	A total *n* = 49 participants (30.6%) dropped out after starting treatment, including *n* = 25 from the PE group (30.9%) and *n* = 24 from the CPT group (30.4%). The final PAI model consisted of four moderator (prescriptive) variables, and two predictor (prognostic) variables. (Non-significant trends were included when they were consistently selected by the bootstrapped variable selection process.) Prescriptive variables: Patients were more likely to dropout from PE, relative to CPT, when they reported higher current relationship abuse (log odds = −1.08, 95% CI [−2.20, −0.13], SE = 0.52, *p* = .037); belonging to a racial minority (log odds = 1.96, 95% CI [0.17, 3.88], SE = 0.94, *p* = .037); reported more severe childhood physical abuse (log odds = −0.83, 95% CI [−1.80, 0.07], SE = 0.47, *p* = .078); and higher levels of anger (log odds = −0.90, 95% CI [−1.94, 0.07], SE = 0.51, *p* = .075). Prognostic variables: Patients were more likely to complete either treatment when they scored higher on the quick IQ test (log odds = 0.60, 95% CI [0.10, 1.13], SE = 0.26, *p* = .021); or completed more years of education (log odds = 0.45, 95% CI [−0.05, 0.98], SE = 0.26, *p* = .091). 19.7% of patients who received their model-indicated optimal treatment dropped out, compared to 40.5% of patients received their model-indicated suboptimal treatment (log odds = 1.02, 95% CI [0.32, 1.75], *z* = 2.80, *p* = .005; relative risk of dropout = 0.49, 95% CI [0.29, 0.82], number needed to treat = 4.8, 95% CI [2.9, 14.7]).	Findings of this exploratory analysis suggest that machine-learning and bootstrapping methodologies may be used to effectively predict optimal treatment for each patient, and reduce the likelihood that patients will dropout before benefitting from treatment. Models such as this could potentially inform decision-support tools to be used in clinical practice. Limitations: It is not clear whether this will generalise to other primary trauma populations; several potentially important variables were not included such as biomarkers and patient preference; all selected variables predicted greater probability of dropout in PE, which could suggest that variables important for predicting dropout from CPT were not included in the analysis; patients classified as dropouts may have benefitted from treatment before they dropped out; the model was only tested on the same data used to train it (internal cross-validation); only investigated two commonly used treatments for PTSD, CPT protocol has been updated since the RCT took place and there is evidence that the new protocol may reduce dropout.
Kratzer et al. (2019)	Somatoform symptoms (HEALTH-49), complex dissociative disorder, mindfulness (FMI) (3)	None	52% of the sample (*n* = 78) had a reliable change in symptoms. The conditional inference tree identified three predictors of reliable change: Lower somatoform symptom (HEALTH-49) score, absence of diagnosis of a complex dissociative disorder, and higher	Results consistent with evidence that dissociation negatively affects therapy outcome. The significance of somatoform symptoms has been overlooked thus far.

Study	Predictors in Final Model(s) (n Predictors in Model)	Evaluation Metrics Reported	Findings	Study Author's Interpretation
			mindfulness (FMI) score. In the subgroup of patients who scored above 3.29 for somatoform symptoms (node 7; $n = 29$), the rate of reliable improvement fell to 27.6%. In the subgroup of patients who scored below 3.29 for somatoform symptoms and were diagnosed with a complex dissociative disorder (Node 6; $n = 33$), the rate of reliable improvement was 39.4% ($p < 0.05$). The subgroup of patients who scored below 3.29 for somatoform disorders but were not diagnosed with a complex dissociative disorder (node 3; $n = 88$) could be divided further into subgroups based on their mindfulness scores. In the subgroup of patients who scored below 3.29 for somatoform disorders, were not diagnosed with a complex dissociative disorder, and scored above 52 on the FMI (Node 5; $n = 70$), the rate of reliable improvement was 71.4%. In the subgroup of patients who scored below 3.29 for somatoform disorders, were not diagnosed with a complex dissociative disorder, and scored 52 or below on the FMI (Node 4; $n = 18$), the rate of reliable improvement was 38.9%.	The finding that mindfulness deficits impair PTSD treatment outcome is novel. The role of mindfulness in development, maintenance, and treatment of PTSD is unclear. Limitations: Severity of PTSD and comorbid disorders was not assessed through structured interview; the duration of stay was not included as a covariate, and exact duration of trauma-focussed treatment was not recorded for each patient (this could also be included as a covariate); lack of follow-up measures means that long-term benefit of treatment was not assessed.
López-Castro et al. (2021)	Age, slope of improvement in PTSD (MPSS-SR), years since last traumatic event, baseline PTSD severity (CAPS, MPSS-SR), age at earliest traumatic event, slope of improvement in problem substance use, employment, interaction between age and slope of improvement in PTSD (MPSS-SR) (9) Emotion regulation (DERS) and baseline primary substance use were selected by the random forest, but were omitted from the regression model as they were not present in the replication dataset	None	Age × Weekly Improvement in PTSD Symptoms (MPSS-R) interaction effect was significant in the test set ($\beta = 0.0104$, SE = 0.0037, $p = .01$), and suggests that older patients with more weekly improvement in symptoms were likely to attend more sessions, whereas younger patients with more weekly improvement in symptoms were likely to attend fewer sessions. This effect was also significant in the validation set ($\beta = 0.0050$, SE = 0.0026, $p = .05$). Employment was a significant predictor in the test set ($\beta = 0.213$, SE = 0.0969, $p = .030$), indicating that patients without employment in the past three years were likely to attend more sessions than patients with full-time or part-time jobs. However, this effect was not significant in the validation set ($\beta = 0.014$, SE = 0.1211, $p = .905$). In the validation set, Weekly Improvement in PTSD symptoms (MPSS-SR) was a significant predictor of number of sessions attended ($\beta = -0.2219$, SE = 0.1117, $p = .05$), but this was not a significant predictor in the training set. No other predictors were significant.	Random forest can identify reproducible predictors of PTSD/SUD treatment attendance, including an interaction between predictors. Some variables selected by random forest were not found to be statistically significant predictors in the Poisson regression analysis, potentially because they were highly correlated with age, PTSD symptom improvement, and/or their interaction. Within treatment improvement in symptoms seems to be an important predictor of attendance regardless of treatment type, this is congruent with the only other published study of this variable in PTSD+SUD treatment, however the finding that baseline PTSD did not interact with change in PTSD is not. The finding that baseline employment status was not significant in the validation set reflects broader pattern in the literature of baseline predictors of PTSD+SUD trial

Study	Predictors in Final Model(s) (*n* Predictors in Model)	Evaluation Metrics Reported	Findings	Study Author's Interpretation
				attendance failing to replicate. The finding that age interacted with weekly improvement adds depth to previous findings that older age is associated with greater session attendance in clinical trials of treatment for PTSD+SUD. Limitations: Analysis only included those who attended at least one session, therefore this study does not generalise to patients who drop out before starting treatment; limited by sample size and measures of the RCT datasets; the validation dataset was missing two of the candidate predictors of the training dataset; many variables identified by literature review as potentially related to attendance were missing from both datasets, (e.g., income, cognitive functioning, anxiety sensitivity, comorbid personality disorder, therapeutic alliance).
Nixon et al. (2021)	No pre-treatment variables were associated with outcome	Examined AUC-ROC but did not report statistics, only presented and visually interpreted plots	Participants were classified as follows: Clear responders (*n* = 94) reported a reliable (≥ sample RCI) decrease in PDS score by session 6, and did not meet CAPS PTSD diagnostic criteria at post-treatment or follow-up. Delayed responders (*n* = 52) did not report reliable decrease in PDS score by session 6, but did report a reliable decrease and did not meet CAPS PTSD diagnostic criteria at post-treatment or follow-up. Partial responders (*n* = 17) reported a reliable decrease in PDS score at post-treatment or follow-up but still met CAPS PTSD diagnostic criteria. Non-responders (*n* = 16) did not report a reliable decrease in PDS score and still met the CAPS PTSD diagnostic criteria at post-treatment or follow-up. The classes were not distinguishable using pre-treatment predictor data.	Results of this exploratory study suggest that it is not possible to predict response pattern pre-treatment from the predictors included in this study. Session-by-session progress data is more informative. Limitations: RCT data, findings require replication in other contexts and samples; completer only analysis, provides no information about predictors of dropout, or the relationship between dropout and non-response; potentially important predictors missing from the dataset such as treatment credibility ratings, homework compliance, and therapeutic alliance; used slightly different versions of self-report PTSD and depression measures (PDS used in 3/4 studies and PSS used in the other, BDI-II used in 3/4 studies and BDI used in the other); the best statistical method for analyses such as these is as yet unknown.
Stirman et al. (2021)	Clinician-rated PTSD symptom severity (CAPS), military sexual trauma (MSIW), physical functioning (PCS), mental functioning (MCS),	R^2 and RMSE with 10 fold cross validation, mean averaged over 1000 runs:	In the final regression model higher baseline PTSD symptoms (β = 0.46, SE = 1.37) and experience of military sexual trauma (β = 0.12, SE = 2.86) were associated with higher posttreatment PTSD symptoms (CAPS), whereas better physical (β = − 0.23, SE = 1.29)	Although on average there was a slight advantage of PE over PCT, for patients with the best overall prognosis PE was associated with substantially better outcomes, but for the patients with poorer prognosis there was no difference

Study	Predictors in Final Model(s) (*n* Predictors in Model)	Evaluation Metrics Reported	Findings	Study Author's Interpretation
	perceived treatment credibility (CEQ) (5)	R^2 = 0.39, 95% CI [0.38587, 0.38789] RMSE = 20.28, 95% CI [20.27, 20.28]	and mental functioning (β = − 0.17, SE = 1.41) and higher perception of treatment credibility (β = − 0.17, SE = 1.24) were associated with lower post-treatment PTSD symptoms (CAPS). These variables were combined to form the PI. Regression model predicting post-treatment CAPS from Treatment Type, the PI, and the Treatment Type by PI interaction explained 39% of the variance in post-treatment PTSD severity. The interaction term was significant (β = 0.2999, SE = 0.1526, 95% CI [0.0008, 0.59906], *t* = 1.97, *p* = 0.0494), as well as the effect of Treatment Type (β = −6.7088, SE = 2.4854, 95% CI [−11.580, −1.8375], *t* = −2.70, *p* = 0.0069), and the PI (β = 0.9154, SE = 0.0765, 95% CI [0.7654, 1.06547], *t* = 11.96, *p* < .0001), indicating that the interaction between the PI and treatment type moderated treatment outcome. For the 64% of patients who were predicted the best treatment outcomes, PE was associated with statistically significantly better outcomes than PCT. Whereas for the 36% of patients with the worst predicted treatment outcomes, treatment type was not associated with outcome. PCT was not associated with better treatment outcomes at any point on the PI continuum.	in outcome between PE and PCT. Replication is necessary before drawing conclusions. Findings suggest that PI may inform treatment selection by identifying patients for whom trauma focussed interventions confer a significant advantage. Limitations: Variable selection and imputation was performed using the whole sample, risking invalid statistical inference, model overfitting, inflated relationships, and overconfidence; Luedtke et al. (2019) suggest *n* = 300 per treatment arm to detect reliable improvements in outcomes related to treatment selection models, RCTs with samples this size currently unavailable for PTSD, sample size too small to facilitate holdout sample, external validation required; source RCT not designed to inform treatment selection; sample all female, mostly veterans, with high chronicity and diagnostic complexity; PE therapists mostly inexperienced, further research with more diverse patient and therapist characteristics recommended, model may not generalise beyond female military population; other predictor variables (e.g., trauma history details) and outcome variables (e.g., quality of life, functioning) could be included.
Stuke et al. (2021)	Posttraumatic cognitions (PTCI), centrality of trauma event to person's identity and life story (CES-7), depression (BDI), gender, general psychopathology (BSI), PTSD symptoms (DTS, PDS), comorbid affective disorder, psychosocial functioning (IMET), rumination (PTQ), age, comorbid substance use disorder (12)	Linear regression: *R* = 0.214, *p* = .021 ADAboost regressor: *R* = 0.162, *p* = .081	Univariate correlations indicated that severe posttraumatic cognitions (PTCI; *r* = .277), greater centrality of traumatic event to patients' identity and life story (CES-7; *r* = .202), and more severe depression (BDI; *r* = .201) were associated with poorer outcome.	Treatment outcome could be significantly predicted from pre-treatment total scores on psychometric measures using linear regression. The overall predictive power of the model was low compared to similar studies using machine learning methods, and arguably too low to be clinically useful. No single predictor was particularly strong, but a higher level of posttraumatic cognitions was the strongest predictor of poor outcome. This finding is in line with one previous study but contradicts numerous others that found no effect of posttraumatic cognitions. The finding that

Study	Predictors in Final Model(s) (*n* Predictors in Model)	Evaluation Metrics Reported	Findings	Study Author's Interpretation
				higher baseline PTSD predicted greater improvement in symptoms is in-line with some previous studies but contradicts others that found the opposite effect (could be due to ceiling effect or over-reporting of baseline symptoms). There is a need to identify more powerful predictor variables and include interactions. Study strengthened by cross-validation methods and naturalistic design. Limitations: Routinely collected data, self-report outcome, lack of control group, lack of formal control for therapy adherence, may have omitted important predictor variables (e.g., employment and social problems), small sample size, large proportion of participants excluded for missing values or dropout, cannot rule out bias due to completer only analysis.
Zhang et al. (2021)	Resting state EEG/PEC features primarily selected from the beta frequency band and eyes-open condition (NR)	None	In the first sample, cluster analysis identified two subtypes in the subsample that met diagnostic criteria for PTSD (*n* = 106), and the stability analysis (100 repetitions on random 90% subsamples) confirmed this as the most stable solution. The clustering stability was significantly lower and more variable in the subsample of healthy controls (*n* = 95). The two subtypes primarily differed in PEC patterns in regions located in the frontoparietal control network and the default mode network; compared with subtype 2, subtype 1 PEC was stronger between the frontal cortex and other regions but weaker between the parietal cortex and other regions. There were no significant differences in clinical or demographic variables between the two subtypes. The cluster analysis was successfully replicated in the second sample. Again, there was no significant difference in clinical or demographic variables between subtype 1 and 2. In the first cohort (*n* = 72), subtype 1 had significantly better treatment outcomes than subtype 2 (group x time interaction: $F(1,123) = 9.04$, $p = 0.0032$, Cohen's $d = 0.80$ for CAPS-IV; $F(1,123) = 4.38$, $p = 0.039$, Cohen's	Subtype 1 represents a subgroup of patients who meet CAPS/CAPS-5 diagnostic criteria for PTSD but do not differ biologically from healthy controls. This is consistent with previous psychiatric neuroimaging findings. However, it may be that patients in subtype 1 differ from healthy controls on other neurological variables that were not measured in this study. Unlike previous studies, no *a priori* assumptions were made about brain regions or frequency bands of interest, leading to unbiased, data-driven, identification of significant biomarkers, which generalised across independent datasets. The findings of this study support the findings of previous fMRI studies. Limitations: Analysis of clusters as predictors of treatment response require replication; EEG vulnerable to confounding neural signals and artifacts of volume conduction; possible change in neural connectivity over time was not accounted for in this study.

Study	Predictors in Final Model(s) (n Predictors in Model)	Evaluation Metrics Reported	Findings	Study Author's Interpretation
			$d = 0.59$ for CAPS-5). Splitting the cohort by treatment, subtype 1 had significantly better outcomes for both PE ($F(1,38) = 7.23$, $p = 0.011$ for CAPS-IV; $F(1,38) = 2.90$, $p = 0.097$ for CAPS-5) and CPT ($F(1,81) = 4.75$, $p = 0.032$ for CAPS-IV; $F(1,81)=2.41$, $p = 0.12$ for CAPS-5). This was replicated in the second cohort ($n = 63$): Subtype 1 had significantly better treatment outcomes than subtype 2 (group × time interaction: $F(1,109) = 4.76$, $p = 0.031$, Cohen's $d = 0.56$ for CAPS-IV; $F(1,109) = 4.46$, $p = 0.037$, Cohen's $d = 0.55$ for CAPS-5). Splitting the cohort by treatment, Subtype 1 had significantly better outcomes for both PE ($F(1,34) = 3.09$, $p = 0.088$ for CAPS-IV; $F(1,34) = 9.31$, $p = 0.0044$ for CAPS-5) and CPT ($F(1,71)=2.13$, $p = 0.15$ for CAPS-IV and $F(1,71) = 0.74$, $p = 0.39$ for CAPS-5). Comparing the percentage of treatment responders, there were significantly more treatment responders in subtype 1 than subtype 2 ($X^2 = 4.07$, $p = 0.044$, number needed to treat = 5.1 for CAPS-IV). The clustering analysis was repeated using only clinical and demographic variables, and the subtypes identified using PEC data could not be identified using clinical and demographic data alone.	

Zhutovsky et al. (2019)	A network centred around the pre-supplementary motor area
	(NR)
		Balanced accuracy = 81.4%
		Sensitivity = 84.8%
		Specificity = 78%
		AUC-ROC (SD) = 0.929 (SD = 0.149)
			Of the $n = 44$ participants treated for PTSD, $n = 24$ met the criteria for treatment response and $n = 20$ did not. The univariate group analysis indicated heightened connectivity in the frontal polar area in non-responders, particularly in the right superior frontal gyrus (p FWE = 0.04).
			The multivariate Gaussian process classification analysis identified a network centred around the pre-supplementary motor area that could be used to classify responders and non-responders with a high degree of accuracy. After Bonferroni correction was applied, no other areas were identified as significant predictors, including the frontal polar area, however this area was significant before applying the Bonferroni correction.
				The results demonstrate that it is feasible to predict individual response to PTSD treatment using resting state fMRI data, this provides a proof-of-concept that PTSD treatment can be personalised using biomarker predictors. Considered alongside previous research into depression treatment response, the results suggest that pre-supplementary motor area connectivity may influence response to treatment regardless of treatment type or disorder.
				The network identified by this study is not comparable to the ventral attention network investigated by Etkin et al. (2019). Although also using resting state fMRI, Etkin et al. (2019) discovered the ventral attention network by comparing patients with PTSD to healthy controls, rather than comparing responders to non-responders, and did not investigate any networks other than ventral attention

Study	Predictors in Final Model(s) (*n* Predictors in Model)	Evaluation Metrics Reported	Findings	Study Author's Interpretation
				network. Therefore, the predictive power of both the pre-supplementary motor area and ventral attention network still require replication. However, both studies demonstrate that resting state fMRI may be used to predict individual treatment response in PTSD. Findings contrast with previous research, this may be due to methodological differences such as previous studies investigating a pre-defined region of interest (as opposed to the data-driven, whole brain approach taken here), previous studies primarily focussed on task-induced changes rather than resting state fMRI, different PTSD populations and types of trauma, different treatment and outcome criteria, different study designs and methods of measurement/analysis. Limitations: Sample size small for machine learning methods; all male veteran sample limits generalizability; mix of treatments means that results are not specific to one particular treatment, and patient by treatment interactions may be obscured.
Zilcha-Mano et al. (2020)	Within-network connectivity in the Executive Control Network (Lateral Prefrontal Cortex right – Posterior Parietal Cortex right; Frontal Pole right – Lateral Prefrontal Cortex right) (NR)	None	Support vector machine identified 18 functional connectivity features that distinguished participants with PTSD (with or without depression) from trauma-exposed healthy controls. Participants with PTSD (with or without depression) displayed lower connectivity in the Within-Executive Control, Within-Salience, Salience–Dorsal Attention, Salience–Default Mode, and Default Mode–Executive Control Networks, and higher connectivity in the Default Mode–Dorsal Attention and Salience–Default Mode Networks (compared to controls). Support vector machine identified 20 functional connectivity features that distinguished participants with PTSD from patients with PTSD+Depression. For those with PTSD alone, within-network connectivity was higher in the Basal Ganglia Network, but lower in the Executive Control, Salience, and Dorsal Attention Networks. The model could classify participants with or without PTSD with an accuracy of 70.6%, and classify those with PTSD with or without	The model identified baseline differences in functional connectivity between patients with PTSD (with or without concurrent depression) and healthy controls, which were significantly associated with change in symptoms over the course PE for PTSD and are therefore clinically useful. The finding that differences in functional connectivity in the identified networks distinguish patients with PTSD from trauma-exposed healthy controls is in-line with some previous findings but not others. The finding that the biomarkers distinguishing PTSD alone from PTSD+Depression did not significantly correlate with treatment response may be due to fact that the treatment was PTSD focussed, and not a treatment for depression. Limitations: Combined data from three trials with different exclusion criteria and different MRI scanners;

Study	Predictors in Final Model(s) (*n* Predictors in Model)	Evaluation Metrics Reported	Findings	Study Author's Interpretation
			MDD with an accuracy of 76.7%. Of the functional connectivity features that distinguished participants with PTSD (with or without MDD) from trauma-exposed healthy controls, there was a significant, positive correlation between within-network connectivity in the Executive Control Network and reduction in PTSD symptoms over the course of PE (Lateral Prefrontal Cortex right–Posterior Parietal Cortex right: *r* = .455, *p* < .001; Frontal Pole right-Lateral Prefrontal Cortex right: *r* = .415, *p* = .002). Such that participants who showed higher levels of functional connectivity in the Executive Control Network pre-treatment reported a greater reduction in PTSD symptoms post-treatment. No significant correlations with change in depression were observed, and there were no significant correlations between the biomarkers identified by the PTSD vs. PTSD+Depression classification and change in PTSD or depression symptoms.	focus on difference between disorders ignored heterogeneity within disorders and transdiagnostic processes; assumed that PTSD and PTSD+Depression are distinct subpopulations, unsupervised ML may be better suited to investigate this; sample size prohibited exploration of symptom cluster associations with functional connectivity; lack of non-trauma-exposed control group prevented exploration of the effect of trauma exposure.

Note. AUC-PR = Area Under the Precision-Recall Curve; AUC-ROC = Area Under the Receiver Operating Characteristic Curve; AUDIT = Alcohol Use Disorder Identification Test; BDI = Beck Depression Inventory; BDI-II = Beck Depression Inventory – II; BSI = Brief Symptom Inventory; CAPS = Clinician-Administered PTSD Scale; CAPS-5 = Clinician-Administered PTSD Scale for DSM-5; CEQ = Credibility/Expectancy Questionnaire; CES-7 = Centrality of Event Scale; CHAID = Chi-square Automatic Interaction Detection; CI = Confidence Interval; CPT = Cognitive Processing Therapy; CTQ = Childhood Trauma Questionnaire; DERS = Difficulties in Emotion Regulation Scale; DSM-5 = Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition; DSM-IV-TR = Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision; DTS = Davidson Trauma Scale; EEG = Electroencephalography; EMDR = Eye Movement Desensitisation and Reprocessing; EQ-5D-5L = EuroQoL 5 Dimensions 5 Levels; FMI = Freiburg Mindfulness Inventory; fMRI = Functional Magnetic Resonance Imaging; FWE = Family Wise Error; HADS = Hospital Anxiety and Depression Scale; HEALTH-49 = Hamburg Modules for the Assessment of Psychosocial Health; IES-R = Impact of Event Scale – Revised; IMET = Index zur Messung von Einschränkungen der Teilhabe; IPE = Intensified Prolonged Exposure; IQ = Intelligence Quotient; MAE = Mean Absolute Error; MINI = Mini International Neuropsychiatric Interview; MMPC LR = Logistic Regression with Max-Min Parent-Child variable selection; MMPI-2 = Minnesota Multiphasic Personality Inventory-2; MOS = Medical Outcome Study; MPSS-SR = Modified Post-traumatic Stress Disorder Symptom Scale Self-Report; MSIW = Military Stress Inventory for Women; NR = Not Reported; OR = Odds Ratio; PAI = Personalised Advantage Index; PCL-5 = PTSD Checklist for DSM-5; PCS = Physical Component Summary Scale; PDS = Posttraumatic Stress Diagnostic Scale; PE = Prolonged Exposure; PEC = Power Envelope Connectivity; PHQ-15 = Patient Health Questionnaire – 15; PHQ-9 = Patient Health Questionnaire-9; PI = Prognostic Index; PSS = Post-traumatic Symptoms Scale; PTCI = Posttraumatic Cognitions Inventory; PTQ = Perseverative Thinking Questionnaire; PTSD = Post-Traumatic Stress Disorder; RCT = Randomised Controlled Trial; RMSE = Root Mean Square Error; SD = Standard Deviation; SE = Standard Error; STAIR+PE = Skills Training in Affective and Interpersonal Regulation plus Prolonged Exposure; SUD = Substance Use Disorder; Tf-CBT = Trauma-focussed Cognitive Behavioural Therapy; URICA = University of Rhode Island Change Assessment; WSAS = Work and Social Adjustment Scale.

APPENDIX F

Predictor Variables in Studies Included in Systematic Review

APPENDIX F - Table 1

Predictor Variables

Study	Candidate Predictor Variables	*N* variables	Predictor Selection Method	Predictors in Final Model(s) (*n* Predictors in Model)
Deisenhofer et al. (2018)	Depression (PHQ-9) Functional impairment (WSAS) Anxiety (GAD-7) Long-term medical condition (yes / no) Disability (yes / no) Antidepressant medication (prescribed / not prescribed) Gender (male / female) Age Employment status (employed, student / unemployed, long-term sick, other)	9	Genetic algorithm	*Tf-CBT:* Functional impairment (WSAS) Employment status Age Gender (4) *EMDR:* Depression (PHQ-9) Medication (2)
Etkin et al. (2019)	MRI EEG Sustained attention (task) Working memory (task) Verbal memory (task) Inhibitory control (task) Response inhibition (task) Flexibility (task) Processing speed (task)	Unclear	Generalised linear modelling	Verbal memory delayed recall impairment Within ventral attention network functional connectivity (2)
Fleming et al. (2018)	Marital Status (single / married / divorced / widow(er) / separated engaged) Military Branch (Army / Navy / Marines / Air Force / National Guard / multiple branches) Active Duty (yes / no) Conflict (OIF / OEF / both) Prior treatment for PTSD (yes / no)	51	Exhaustive CHAID	Days between introductory session and therapy invitation URICA item 9 ("*I have been successful in working on my trauma issues, but I'm not sure I can keep up the effort on my own*") URICA item 5 ("*I'm not the one with trauma issues, it doesn't make much sense for me to be here*") Presence of traumatic brain injury

	Number of children Number of combat tours Years of education Traumatic brain injury severity (none / mild / moderate or severe) Depression (yes / no) Anxiety disorder (yes / no) Alcohol use disorder (yes / no) Substance use disorder (yes / no) Psychotropic medication (taking / not taking) Trauma type (combat only / sexual only / both) Pre-military trauma (yes / no) Pending disability claim (yes / no) Total VA Disability Rating (10% or below / 20-40% / 50-70% / ≥ 80%) PTSD Disability Rating (10% or below / 20-40% / 50-70% / ≥ 80%) PTSD symptom severity (PCL total, subscales, and items) Days between mental health consult and mental health appointment Days from psychotherapy referral to introduction session Days from completed introduction session to therapy invitation Motivational readiness to address trauma (URICA total, subscales, and items) PTSD Information Session Survey (total and items [*n* = 9]) Format of introduction session contact (group / individual) Weighs pros or cons more (pros / cons / equal weight) Readiness category (precontemplation / contemplation / preparation) Treatment choice (PE / CPT group / CPT individual / CPT individual or group / other / none)			Prior PTSD treatment (5)
Forbes et al. (2003)	Psychopathology (MMPI-2 clinical subscales [*n* = 10] validity subscales [*n* = 3] and factors [*n* = 3]) PTSD (PCL) Anxiety (HADS) Depression (HADS) Alcohol use (AUDIT)	21	Ward's hierarchical cluster analysis	Clinical subscales and validity subscales of MMPI-2 anxiety and depression (HADS) Alcohol use (AUDIT) (16)

	Combat exposure (CES)			
Held et al. (2022)	Age Gender (male / female) Race (white / all other races) Ethnicity (non-Hispanic / Hispanic or Latino) Education (8 categories dummy coded) Marital Status (married or domestic partnership / not married) Military Branch (5 categories dummy coded) Deployed (yes / no) Served in military after 11th September 2001 (yes / no) Military sexual trauma (three items measured yes / no) Referral source (6 categories dummy coded) Clinician assessed PTSD symptoms (CAPS-5) Self-reported PTSD symptoms (PCL-5) Depression (PHQ-9) Alcohol use disorder (AUDIT-C) Neurobehavioral symptom exaggeration (NSI-Valid) Trauma related cognitions (PTCI)	104	Elastic net classification Gradient boosted models Random forest Ridge classification Logistic regression Logistic regression with Max-Min Parent-Child variable selection	*Gradient Boosted Models predicting Fast Response class:* Post-traumatic stress symptoms (item level responses to PCL-5 and CAPS-5) Post-traumatic cognitions (item level responses to PTCI) (10) *Elastic Net predicting Minimal Response class:* Post-traumatic stress symptoms (item level responses to PCL-5 and CAPS-5) Post-traumatic cognitions (item level responses to PTCI) Marriage/domestic partnership status Age Gender Level of education Military service branch (60)
Hendriks et al. (2018)	Age Educational level Living alone PTSD symptom severity (PSS-SR) Depression (BDI-II) Dissociative symptom severity (DES) Borderline personality disorder (BPD-47) Psychoactive medication use Fear activation during first exposure session Within-session fear habituation during the first session Between session fear habituation	14	Stepwise multinominal logistic regression	Living condition Between-session fear habituation (2)
Herzog et al. (2021)	Age Gender Number of children Inability to work Wish to retire School-leaving qualification Professional qualification Occupational status (6 categories dummy coded)	≥ 46 (unclear)	Elastic net	PTSD severity (IES-R total) Psychoticism (BSI) Avoidance (IES-R subscale) Wish to retire Depression (BDI-II) Number of comorbid diagnoses Age Bronchial asthma

	Marital status (4 categories dummy coded) In a relationship (yes / no) Living situation (7 categories dummy coded) Previous outpatient psychiatric care Previous outpatient psychotherapy ICD-10 medical diagnoses (dummy coded, including but not limited to: Bronchial Asthma; Endocrine, nutritional and metabolic disease; Personality Disorder; Obesity) Number of comorbid diagnoses PTSD symptoms (IES-R) Psychiatric symptoms and distress (BSI, 9 subscales) Life satisfaction (SWLS) Depression symptoms (BDI-II) Depression, anxiety, and somatic symptoms (PHQ, 3 scales) Psychosocial functioning (GAF) Work disability last year Duration of work disability			Somatic symptoms (PHQ-15) Outpatient psychiatric care Children Being retired Work disability in past year Outpatient psychotherapy Somatization (BSI) (15)
Hoeboer et al. (2021)	Patient expectancies (expectancy of burden and credibility questionnaire) Age Gender Cultural background Education Employment Social support (MOS) Childhood trauma background (CTQ) General health status (EQ-5D-5L) Depression (BDI) Post-traumatic cognitions (PTCI) Interpersonal problems (IIP) Self-esteem (RSES) Emotion regulation difficulties (DERS) Somatoform dissociation (SDQ-5) Presence of personality disorder (SCID-2) Number of DSM-IV-defined Axis-1 disorders excluding PTSD (MINI) Dissociation (DSP-I)	24	Boruta algorithm	*CAPS-5 change following PE/IPE:* Depression (BDI) Social support (MOS) Concurrent mental health problems (MINI axis 1) Childhood sexual abuse (CTQ) (4) *CAPS-5 change following STAIR+PE:* General health status (EQ-5D-5L) Emotion regulation difficulties (DERS) PTSD severity (CAPS-5) (3) *PCL-5 change following PE or IPE:* Depression (BDI) Social support (MOS) (2) *PCL-5 change following STAIR and PE:* General health status (EQ-5D-5L)

	PTSD symptom severity (CAPS-5) Psychotropic medication			Emotion regulation difficulties (DERS) (2)
Keefe et al. (2018)	Age Race Years of education Estimated IQ Years since index rape Severity of childhood physical abuse Severity of childhood sexual abuse Abuse by current partner Total sex crime exposures CAPS total score PTSD avoidance (PSS) PTSD arousal (PSS) PTSD re-experiencing (PSS) Depression (BDI-II) Dissociation (DES) Hopelessness (BHS) Trait anger (STAXI) Total trauma cognitions (TRGI)	20	Bootstrapped, random forest variant of model-based recursive partitioning, and bootstrapped variant of an AIC-based backward selection model	*Prescriptive variables:* Childhood physical abuse Current relationship abuse Trait anger Race *Prognostic variables:* Years of education Estimated IQ score (6)
Kratzer et al. (2019)	Depression (HEALTH-49) Somatoform symptoms (HEALTH-49) Well-being (HEALTH-49) Presence of a complex dissociative disorder Mindfulness (FMI) *Candidate variables excluded through univariate analysis:* Gender Age Presence of a personality disorder Childhood trauma (CTQ) PTSD symptoms (IES-R) Interactional difficulties (HEALTH-49) Self-efficacy (HEALTH-49) Phobic anxiety (HEALTH-49) Social distress (HEALTH-49) Social support (HEALTH-49)	5	Conditional inference tree	Somatoform symptoms (HEALTH-49) Complex dissociative disorder Mindfulness (FMI) (3)

	Activity and participation (HEALTH-49) Dissociation (DES-T)			
López-Castro et al. (2021)	Age Gender Race and ethnicity Employment Education Marital status Severity of substance use Substance use disorder type (alcohol vs. substance) Comorbid substance use disorder diagnosis PTSD symptom severity (CAPS) Trauma characteristics (more than one traumatic event, number of traumatic events, trauma before age 18, age of first trauma, sexual assault, physical assault, other trauma, accident) Type of intervention (COPE vs. RPT) Depression diagnosis Emotion regulation (DERS) Within-treatment substance use Within treatment PTSD symptom change (MPSS-SR)	28	Random forest	Age Slope of improvement in PTSD (MPSS-SR) Years since last traumatic event Baseline PTSD severity (CAPS, MPSS-SR) Age at earliest traumatic event Slope of improvement in problem substance use Employment Interaction between age and slope of improvement in PTSD (MPSS-SR) (9) Emotion regulation (DERS) and baseline primary substance use were selected by random forest, but were omitted from the regression model as they were not present in the replication dataset
Nixon et al. (2021)	Race Marital status Income level Age Years of education Childhood sexual abuse Childhood physical abuse Endorsed adult sexual violence Adult physical violence Intimate partner violence independent of other forms of violence Index trauma type Depression diagnosis Depression severity (BDI) Panic disorder diagnosis PTSD symptoms (PDS, CAPS)	38	Random forest	No pre-treatment variables were associated with outcome (0)

Stirman et al. (2021)	Clinician-rated PTSD symptom severity (CAPS) Re-experiencing (PCL) Avoidance (PCL) Numbing (PCL) Hyperarousal (PCL) Time since trauma (CAPS) Sexual index trauma (CAPS) Military sexual trauma (MSIW) Military stress exposure (MSIW) Number of trauma types Age White race College Education Married/living as married Working Current mood disorder (SCID-P) Current anxiety disorder (SCID-P) Borderline personality disorder (SCID-P) Other personality disorder (SCID-P) Depression symptoms (BDI) Anxiety symptoms (SSAI) Dissociative symptoms (TSI) Anger symptoms (TSI) Physical Functioning (PCS) Mental Functioning (MCS) Self-reported quality of life (QOLI) Treatment credibility (CEQ) Psychoactive medication use at screening Benzodiazepine use at screening	29	Elastic net with stepwise AIC-penalised bootstrapped variable selection	Clinician-rated PTSD symptom severity (CAPS) Military sexual trauma (MSIW) Physical functioning (PCS) Mental functioning (MCS) Perceived treatment credibility (CEQ) (5)
Stuke et al. (2021)	Age Gender Comorbid ICD-10 psychiatric disorder Trauma details (LEC scale of PDS) PTSD symptoms (DTS, PDS) Posttraumatic cognitions (PTCI) Centrality of trauma event to person's identity and life story (CES-7)	12	Principal components analysis	Posttraumatic cognitions (PTCI) Centrality of trauma event to person's identity and life story (CES-7) Depression (BDI) Gender General psychopathology (BSI) PTSD symptoms (DTS, PDS) Comorbid affective disorder

	Rumination (PTQ) Depression (BDI) General psychopathology (BSI) Psychosocial functioning (IMET)			Psychosocial functioning (IMET) Rumination (PTQ) Age Comorbid substance use disorder (12)
Zhang et al. (2021)	PEC features from eight EEG conditions (four frequency bands across eyes-open/eyes-closed conditions)	Unclear	Sparse k-means clustering	Resting state EEG/PEC features primarily selected from the beta frequency band and eyes-open condition (NR)
Zhutovsky et al. (2019)	Structural MRI and resting state functional MRI data with age and total intracranial volume as covariates	Unclear	Gaussian process classifier	A network centred around the pre-supplementary motor area (NR)
Zilcha-Mano et al. (2020)	43 MRI regions of interest	Unclear	Support vector machine	Within-network connectivity in the Executive Control Network (Lateral Prefrontal Cortex right – Posterior Parietal Cortex right; Frontal Pole right – Lateral Prefrontal Cortex right) (NR)

Note. AIC = Akaike Information Criterion; AUDIT = Alcohol Use Disorder Identification Test; AUDIT-C = Alcohol Use Disorder Identification Test - Consumption; BDI = Beck Depression Inventory; BDI-II = Beck Depression Inventory - II; BHS = Beck Hopelessness Scale; BPD-47 = Borderline Personality Disorder symptom checklist; BSI = Brief Symptom Inventory; CAPS = Clinician-Administered PTSD Scale; CAPS-5 = Clinician-Administered PTSD Scale for DSM-5; CEQ = Credibility/Expectancy Questionnaire; CES = Combat Exposure Scale; CES-7 = Centrality of Event Scale; CHAID = Chi Squared Automatic Interaction Detection; COPE = combined prolonged exposure and relapse prevention therapy; CTQ = Childhood Trauma Questionnaire; DERS = Difficulties in Emotion Regulation Scale; DES = Dissociative Experiences Scale; DES-T = Dissociative Experiences Scale - Taxon; DSP-I = Dissociative subtype of PTSD Interview; DTS = Davidson Trauma Scale; EEG = Electroencephalography; EQ-5D-5L = EuroQoL 5 Dimensions 5 Levels; FMI = Freiburg Mindfulness Inventory; GAD-7 = Generalised Anxiety Disorder 7; GAF = Global Assessment of Functioning; HADS = Hospital Anxiety and Depression Scale; HEALTH-49 = Hamburg Modules for the Assessment of Psychosocial Health; IES-R = Impact of Event Scale - Revised; IIP = Inventory of Interpersonal Problems; IMET = Index zur Messung von Einschränkungen der Teilhabe; LEC = Life Events Checklist; MCS = Mental Component Summary Scale; MINI = Mini International Neuropsychiatric Interview; MMPI-2 = Minnesota Multiphasic Personality Inventory-2; MOS = Medical Outcome Study; MPSS-SR = Modified PTSD Symptom Scale Self-Report; MRI = Magnetic Resonance Imaging; MSIW = Military Stress Inventory for Women; NSI-Valid = Neurobehavioral Symptoms Inventory Validity-10; PCL = PTSD Checklist; PCL-5 = PTSD Checklist for DSM-5; PCS = Physical Component Summary Scale; PDS = Posttraumatic Stress Diagnostic Scale; PEC = Power Envelope Connectivity; PHQ = Patient Health Questionnaire; PHQ-9 = Patient Health Questionnaire-9; PSS = Post-traumatic Symptoms Scale; PSS-SR = PTSD Symptom Scale, Self-Report; PTCI = Posttraumatic Cognitions Inventory; PTQ = Perseverative Thinking Questionnaire; PTSD = Post-traumatic Stress Disorder; QOLI = Quality of Life Inventory; RPT = Relapse Prevention Therapy; RSES = Rosenberg Self-esteem Scale; SCID-2 = Clinical Interview for DSM-IV Personality Disorders; SCID-P = Structured Clinical Interview for DSM-IV Patient Version; SDQ-5 = Somatoform Dissociation Questionnaire ; SSAI = Spielberger State Anxiety Inventory ; STAXI = State-Trait Anger Expression Inventory; SWLS = Satisfaction With Life Scale; TRGI = Trauma-Related Guilt Inventory; TSI = Trauma Symptom Inventory; URICA = University of Rhode Island Change Assessment; WSAS = Work and Social Adjustment Scale.

APPENDIX G

Missing Data in Chapter 3 Dataset

APPENDIX G - Table 1

Percentage of Missing Values on Each Variable Used for Multiple Imputation

	Sample (*N* = 1193)	Tf-CBT (*n* = 1155)	EMDR (*n* = 38)
	% Missing	% Missing	% Missing
Age	0	0	0
Gender	0.3	0.3	0
Disability	0	0	0
LTC	38.6	38.8	34.2
Ethnicity[a]	5.4	5.5	5.3
PHQ-9 pre	3.3	3.4	0
PHQ-9 post	6.2	6.1	7.9
GAD-7 pre	3.4	3.5	0
GAD-7 post	6.1	6.1	7.9
WSAS pre	13.7	13.3	23.7
WSAS post	20.2	20.1	23.7
Employment pre	4.9	5.0	2.6
Medication[b] pre	19.1	19.4	10.5

Note. EMDR = Eye-movement desensitization and reprocessing; GAD-7 = Generalised Anxiety Disorder 7; LTC = Long-term medical condition; PHQ-9 = Patient Health Questionnaire 9; Tf-CBT = Trauma-focussed cognitive behavioural therapy; WSAS = Work and Social Adjustment Scale.

pre = Pre-treatment, first high intensity treatment session.

post = Post-treatment, last high intensity treatment session.

[a] Ethnicity = Office for National Statistics ethnic group.

[b] Medication = Antidepressant medication status.

APPENDIX H

Correlation Matrices and Q-Q Plots for Chapter 3 Dataset

APPENDIX H - Table 1

Spearman's Correlations Between Pre- and Post-treatment Measures for the Tf-CBT Group

			PHQ-9 pre	PHQ-9 post	GAD-7 pre	GAD-7 post	WSAS pre
Spearman's rho	PHQ-9 post	Coefficient	.614				
		Sig. (2-tailed)	<.001				
		N	114				
	GAD-7 pre	Coefficient	.744	.560			
		Sig. (2-tailed)	<.001	<.001			
		N	114	114			
	GAD-7 post	Coefficient	.433	.850	.551		
		Sig. (2-tailed)	<.001	<.001	<.001		
		N	114	114	114		
	WSAS pre	Coefficient	.626	.438	.502	.322	
		Sig. (2-tailed)	<.001	<.001	<.001	<.001	
		N	114	114	114	114	
	WSAS post	Coefficient	.494	.796	.492	.750	.590
		Sig. (2-tailed)	<.001	<.001	<.001	<.001	<.001
		N	114	114	114	114	114

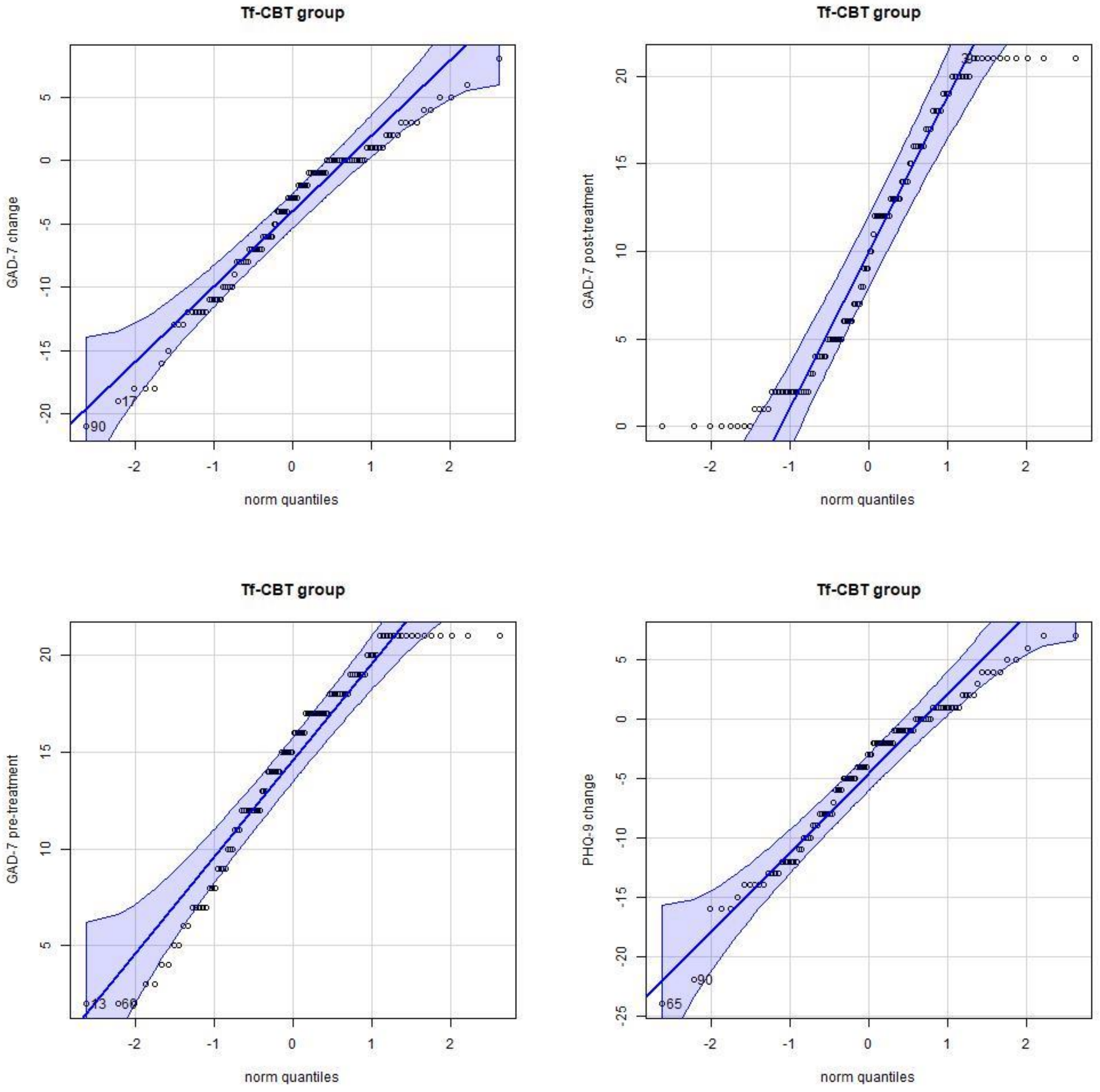
APPENDIX H - Table 2

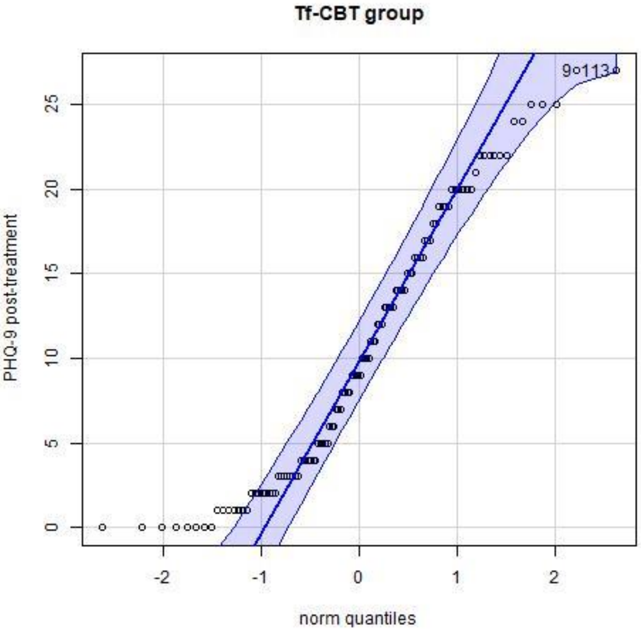
Spearman's Correlations Between Pre- and Post-treatment Measures for the EMDR Group

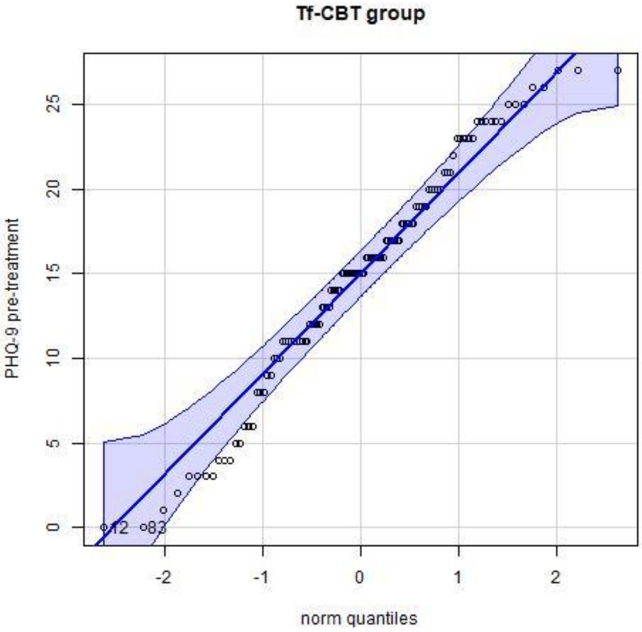
			PHQ-9 pre	PHQ-9 post	GAD-7 pre	GAD-7 post	WSAS pre
Spearman's rho	PHQ-9 post	Coefficient	.678				
		Sig. (2-tailed)	<.001				
		N	38				
	GAD-7 pre	Coefficient	.718	.523			
		Sig. (2-tailed)	<.001	<.001			
		N	38	38			
	GAD-7 post	Coefficient	.588	.842	.571		
		Sig. (2-tailed)	<.001	<.001	<.001		
		N	38	38	38		
	WSAS pre	Coefficient	.689	.491	.412	.430	
		Sig. (2-tailed)	<.001	.002	.010	.007	
		N	38	38	38	38	
	WSAS post	Coefficient	.530	.826	.388	.772	.634
		Sig. (2-tailed)	<.001	<.001	.016	<.001	<.001
		N	38	38	38	38	38

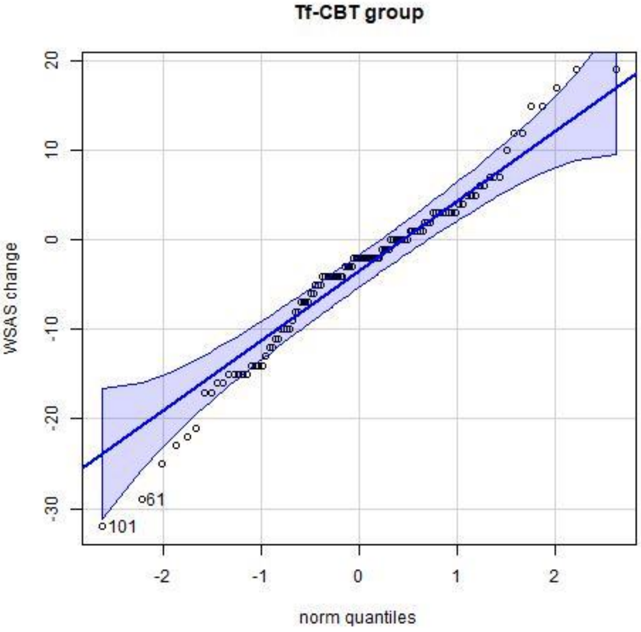
APPENDIX H – Figure 1

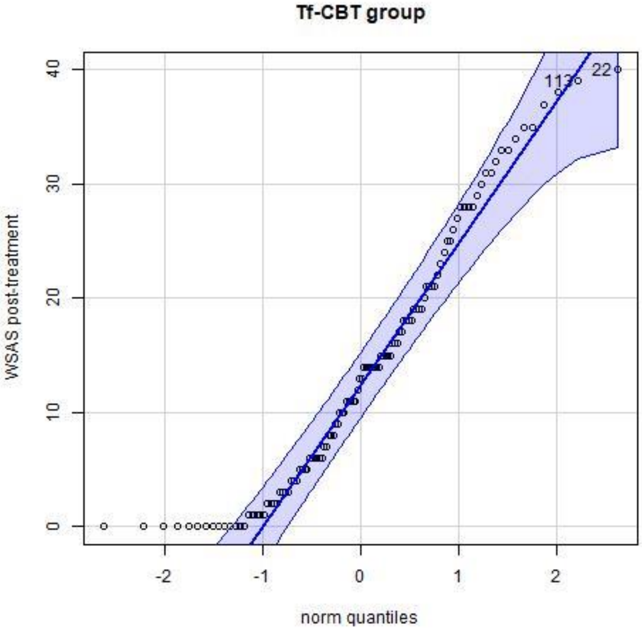
*Q-Q Plots (with 95% Confidence Intervals) Displaying Distribution of Pre-Treatment, Post-Treatment, and Change Scores on GAD-7, PHQ-9, and WSAS for Tf-CBT (*N *= 114) and EMDR (*N *= 38) Groups*

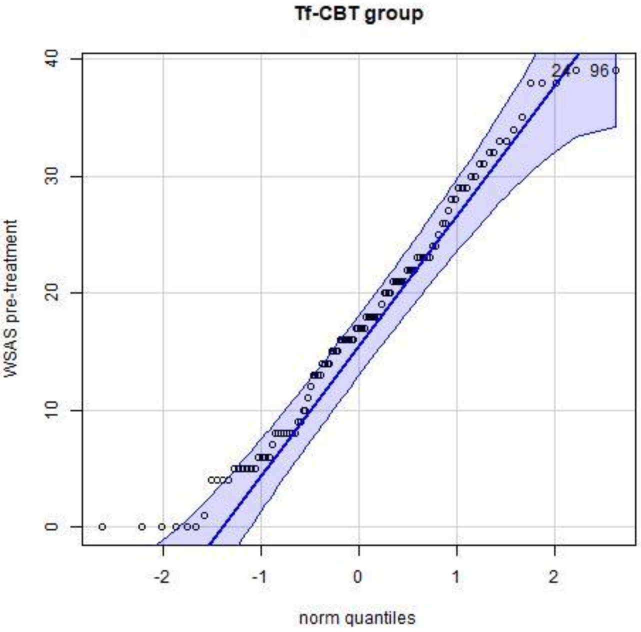


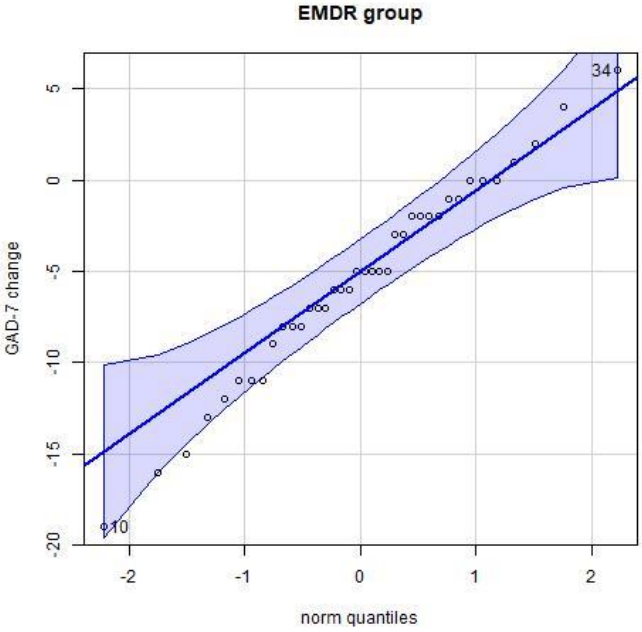

Tf-CBT group
PHQ-9 post-treatment
norm quantiles
0
5
10
15
20
25
-2
-1
0
1
2
9
113

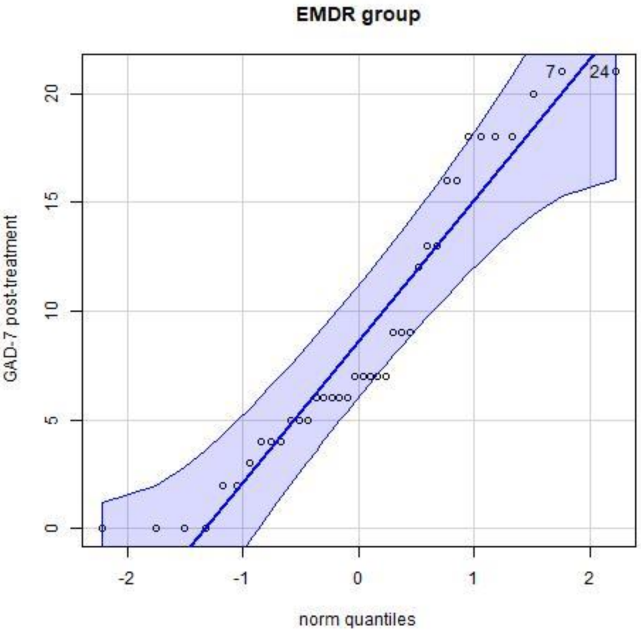

Tf-CBT group
PHQ-9 pre-treatment
norm quantiles
0
5
10
15
20
25
-2
-1
0
1
2
12
83

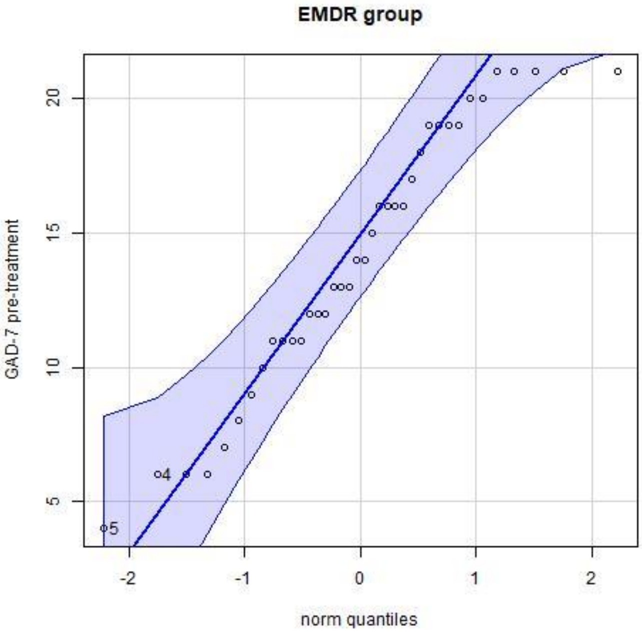

Tf-CBT group
WSAS change
norm quantiles
-30
-20
-10
0
10
20
-2
-1
0
1
2
61
101

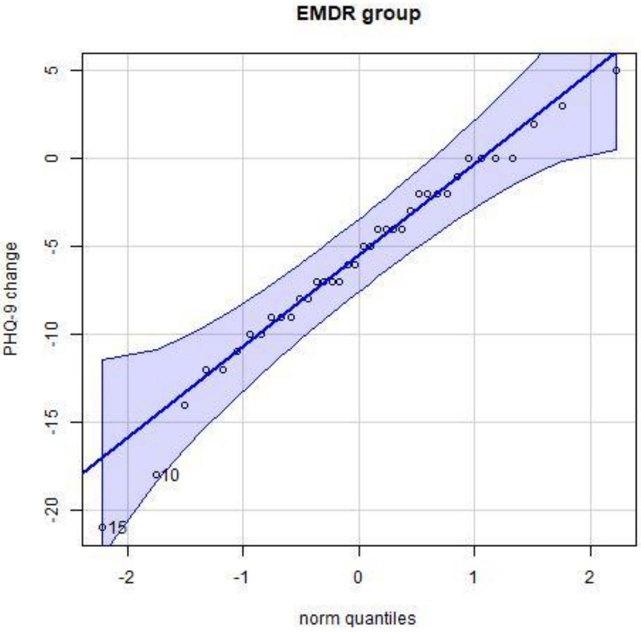

Tf-CBT group
WSAS post-treatment
norm quantiles
0
10
20
30
40
-2
-1
0
1
2
113
22

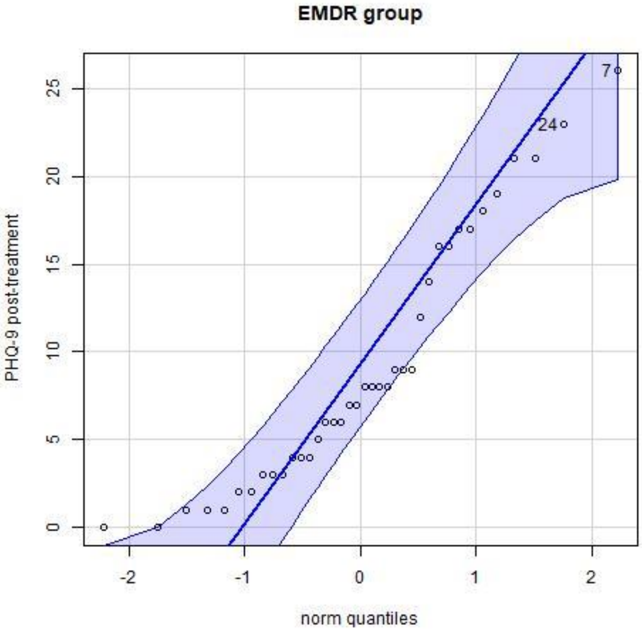

Tf-CBT group
WSAS pre-treatment
norm quantiles
0
10
20
30
40
-2
-1
0
1
2
24
96

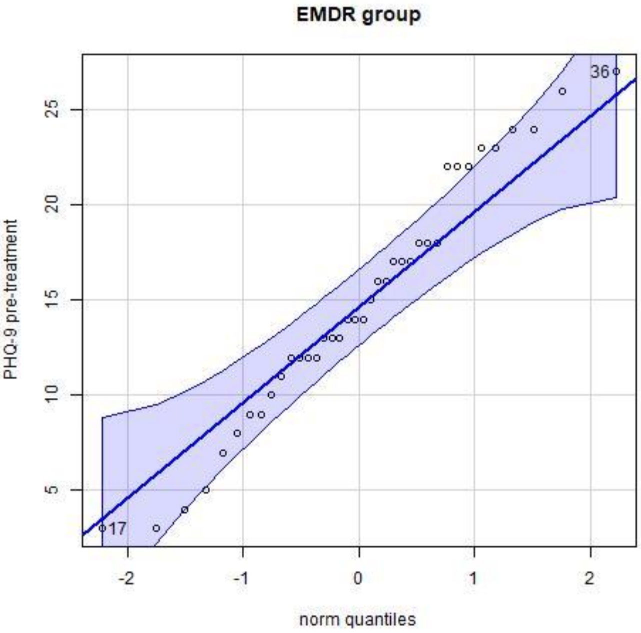

EMDR group
GAD-7 change
norm quantiles
-20
-15
-10
-5
0
5
-2
-1
0
1
2
34
10

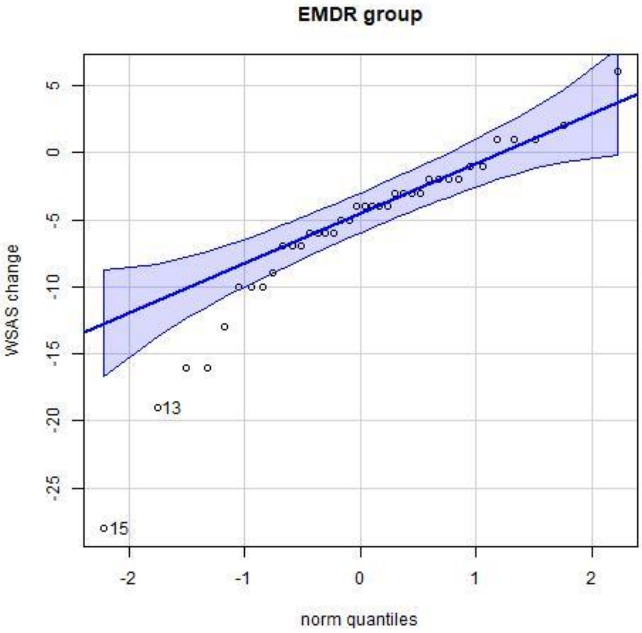

EMDR group
GAD-7 post-treatment
norm quantiles
0
5
10
15
20
-2
-1
0
1
2
7
24

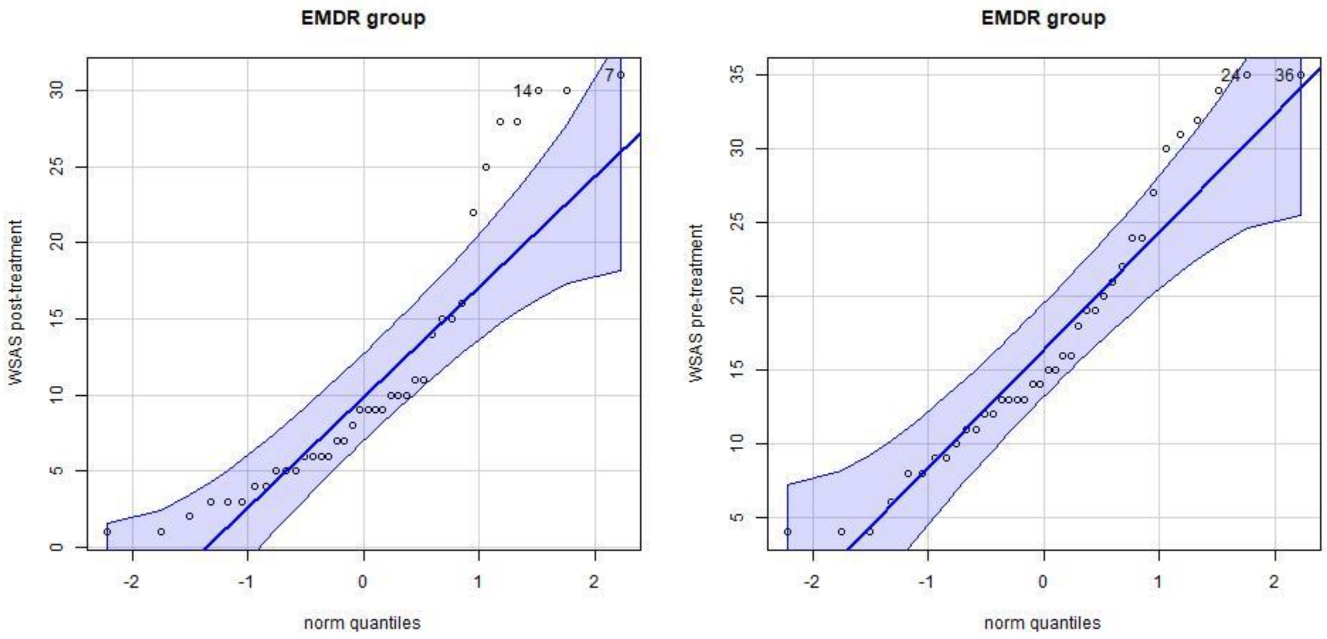

EMDR group
GAD-7 pre-treatment
norm quantiles
5
10
15
20
-2
-1
0
1
2
4
5


EMDR group
PHQ-9 change
5
0
-5
-10
-15
-20
-2
-1
0
1
2
norm quantiles
10
15


EMDR group
PHQ-9 post-treatment
25
20
15
10
5
0
-2
-1
0
1
2
norm quantiles
7
24


EMDR group
PHQ-9 pre-treatment
25
20
15
10
5
-2
-1
0
1
2
norm quantiles
36
17


EMDR group
WSAS change
5
0
-5
-10
-15
-20
-25
-2
-1
0
1
2
norm quantiles
13
15



Note. Q-Q plots created in R, using the `qqPlot` function in the car package.

APPENDIX I

Details of Multiple Regression in Chapter 3

Regression analysis was performed using IBM SPSS Statistics 27. A multiple regression model was tested with post-treatment PHQ-9 score as the dependent variable, and three independent variables: A binary indicator of having received optimal treatment, pre-treatment PHQ-9 score, and propensity score. Patients with a PAI less than the standard deviation of the PAI in the model development sample (1.92) were excluded from this analysis in order to test the clinical utility of the PAI among patients with the most robust model indicated treatment recommendation.

Normality, Homoscedasticity and Linearity

To test the assumptions of normality, homoscedasticity and linearity, the regression model was fitted, and the standardised residuals plotted against the standardised predicted values (APPENDIX I – Figure 1). The standardised residuals scatterplot presented in APPENDIX I – Figure 1 suggested a violation of the assumption of homoscedasticity. As such, the distributions of the dependent variable and continuous independent variables were examined via histograms presented in APPENDIX I – Figure 2.

The histograms presented in APPENDIX I – Figure 2 suggested that post-treatment PHQ-9 score and propensity score were positively skewed. Square root and log transformations were performed for each variable and the resulting distributions examined. Histograms for the square root transformed variables are presented in APPENDIX I – Figure 3, and histograms for the log transformed variables are presented in APPENDIX I – Figure 4.

The square root transformation appeared to reduce the skew for post-treatment PHQ-9 scores, but propensity score still appeared to be positively skewed. The logarithm transformation appeared to reduce the skew for propensity score. The skewness statistics presented in APPENDIX I – Table 1 also suggested that the square root transformation produced the least skewed distribution for pre-treatment PHQ-9 score, and logarithm transformation produced the least skewed distribution for propensity score.

The regression analysis was repeated with square root transformed post-treatment PHQ-9 score as the dependent variable, and the following three independent variables: A binary indicator of optimal treatment, pre-treatment PHQ-9 score, and logarithm transformed propensity score. The standardised residuals scatterplot presented in APPENDIX I – Figure 5 suggested that the transformations had reduced the heteroscedasticity of the residuals.

Screening for Outliers

Case-wise diagnostics found no univariate outliers on any of the variables. Multivariate outliers were investigated via Mahalanobis distance. The critical value for Mahalanobis distance is X^2 $p < .001$ with degrees of freedom equal to the number of independent variables, therefore a Mahalanobis distance greater than 16.266 would indicate a multivariate outlier. The maximum Mahalanobis distance was 7.432, suggesting that there were no multivariate outliers.

Screening for Multicollinearity

APPENDIX I – Table 2 presents the correlations between each of the variables. There were no correlations > .7 between independent variables, suggesting there was no multicollinearity. None of the Tolerance values presented in APPENDIX I – Table 3 are approaching 0, and all the Variance Inflation Factor values were close to 1, suggesting no multicollinearity. The collinearity diagnostics presented in APPENDIX I – Table 4 also suggested an absence of multicollinearity as none of the condition index values are greater than 30.

APPENDIX I - Table 1

Descriptive statistics for variables in the regression analysis

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness	
						Statistic	Std. Error
PHQ-9 pre	93	.00	27.00	13.8602	6.74793	-.207	.250
PHQ-9 post	93	.00	25.00	8.7849	7.39102	.616	.250
PHQ-9 post_sqrt	93	.00	5.00	2.6020	1.42695	-.169	.250
PHQ-9 post_log	93	.00	1.41	.8275	.42022	-.523	.250
propensity_score	93	.01	.22	.0626	.05135	1.451	.250
propensity_sqrt	93	.10	.47	.2326	.09280	.874	.250
propensity_log	93	-1.97	-.65	-1.3312	.33329	.207	.250

APPENDIX I – Table 2

Correlations Between the Variables in the Regression Analysis

		Optimal received	PHQ-9 pre	Propensity score (log)	PHQ-9 post (sqrt)
Optimal received	Pearson Correlation	1	.269**	.113	.278**
	Sig. (2-tailed)		.009	.281	.007
	N	93	93	93	93
PHQ-9 pre	Pearson Correlation	.269**	1	-.171	.649**
	Sig. (2-tailed)	.009		.101	<.001
	N	93	93	93	93
Propensity score (log)	Pearson Correlation	.113	-.171	1	-.124
	Sig. (2-tailed)	.281	.101		.235

APPENDIX I – Table 3

Tolerance and Variance Inflation Factor for each of the Independent Variables

Independent Variable	Tolerance	Variance Inflation Factor
Optimal received	.902	1.109
PHQ-9 pre	.886	1.128
Propensity score (log)	.943	1.060

APPENDIX I – Table 4

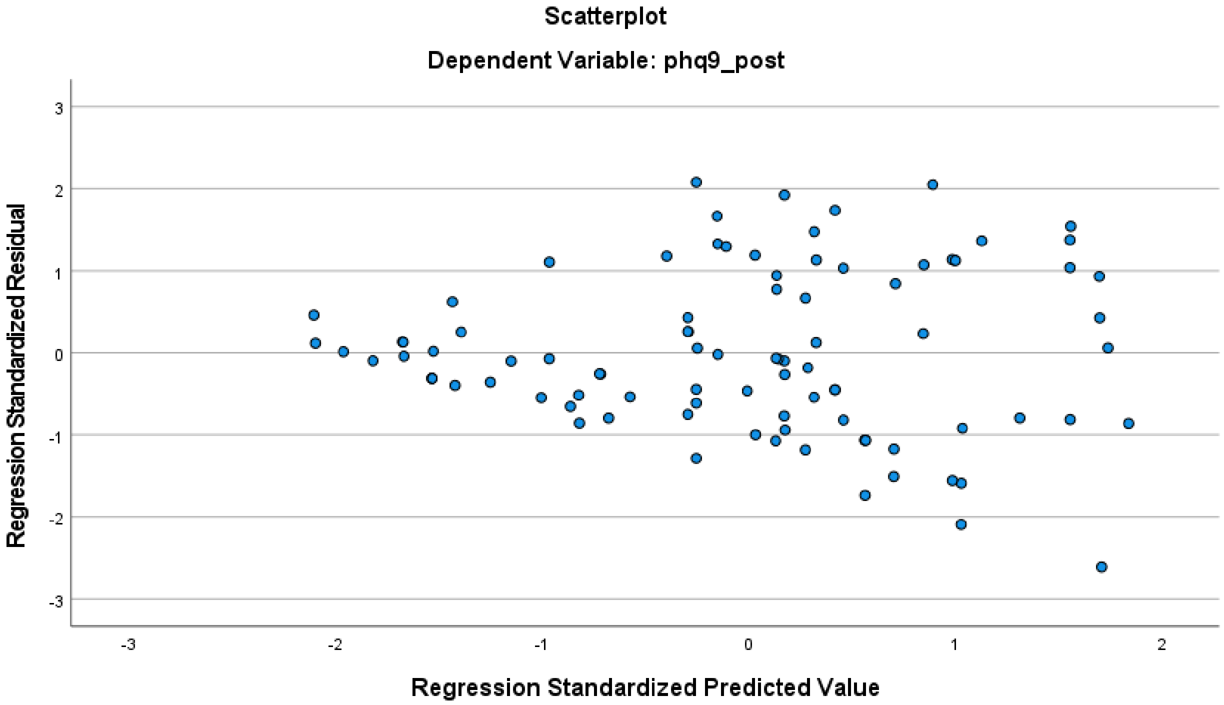
Collinearity Diagnostics

				Variance Proportions			
			Condition		Optimal	PHQ-9	Propensity
Model	Dimension	Eigenvalue	Index	(Constant)	received	pre	score (log)
1	1	3.464	1.000	.00	.02	.01	.00
	2	.384	3.003	.01	.86	.01	.02
	3	.124	5.289	.05	.05	.98	.06
	4	.028	11.151	.94	.06	.00	.91

a. Dependent Variable: PHQ-9 post_sqrt

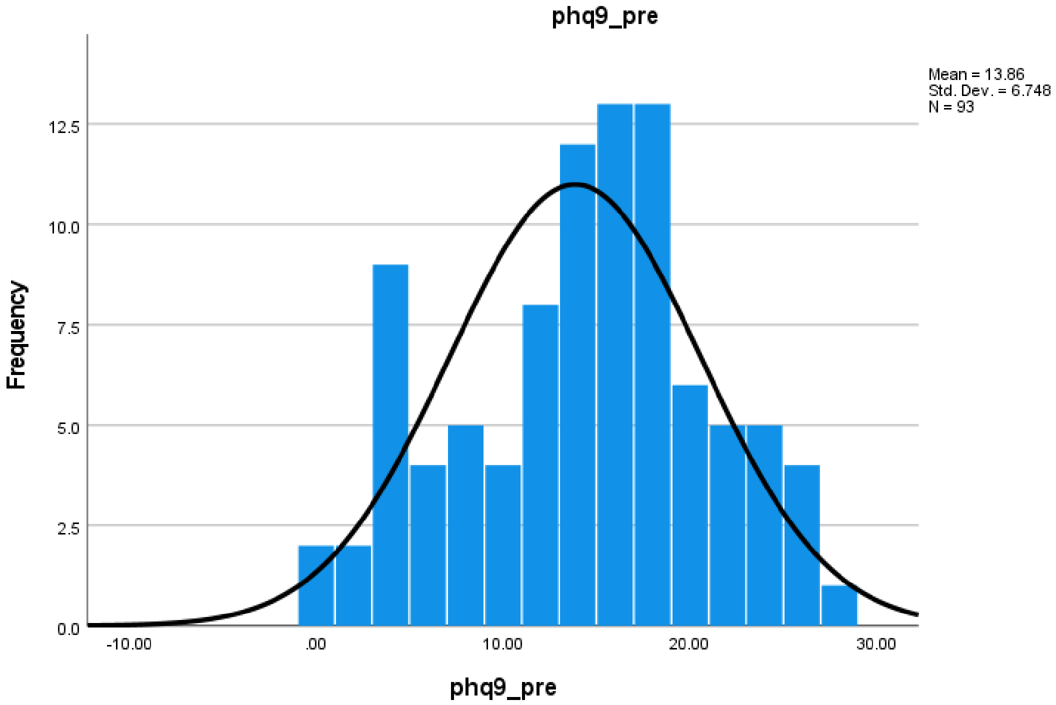
APPENDIX I - Figure 1

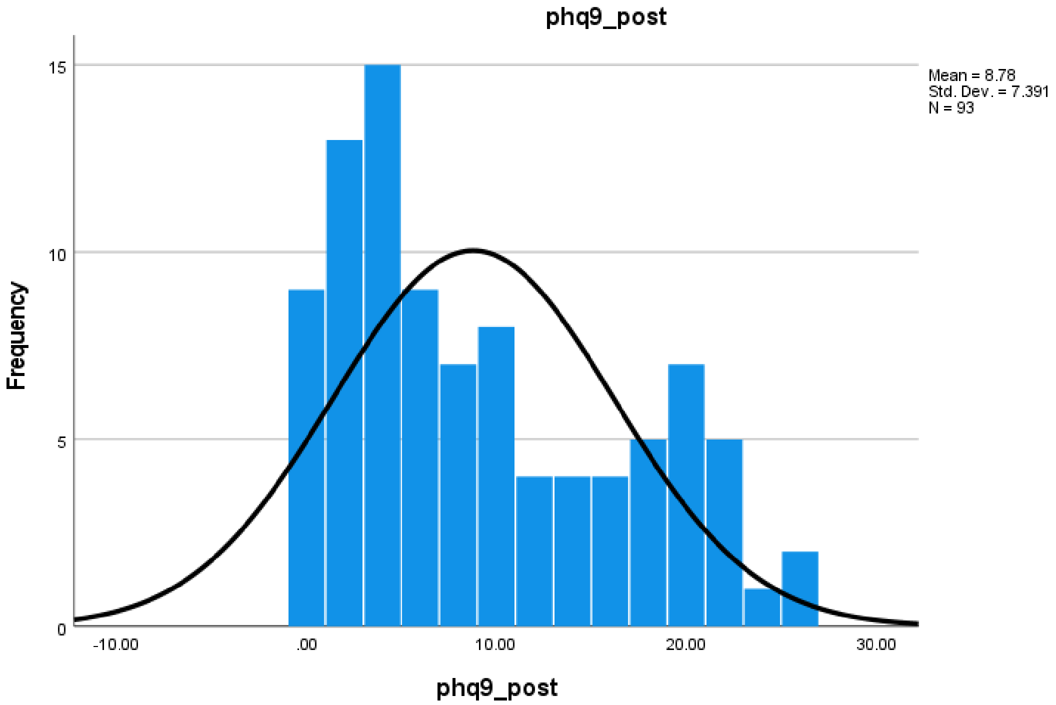
Standardised Residuals Plotted Against Standardised Predicted Post-Treatment PHQ-9 Scores

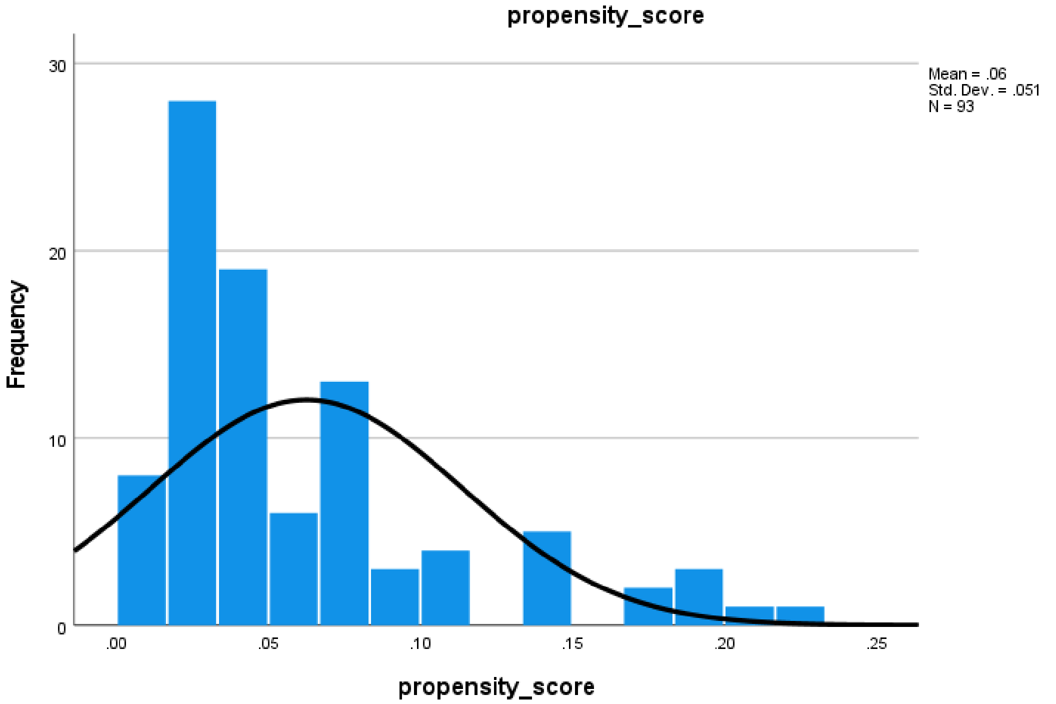


APPENDIX I - Figure 2

Histograms Displaying the Distributions of the Continuous Variables before Transformation

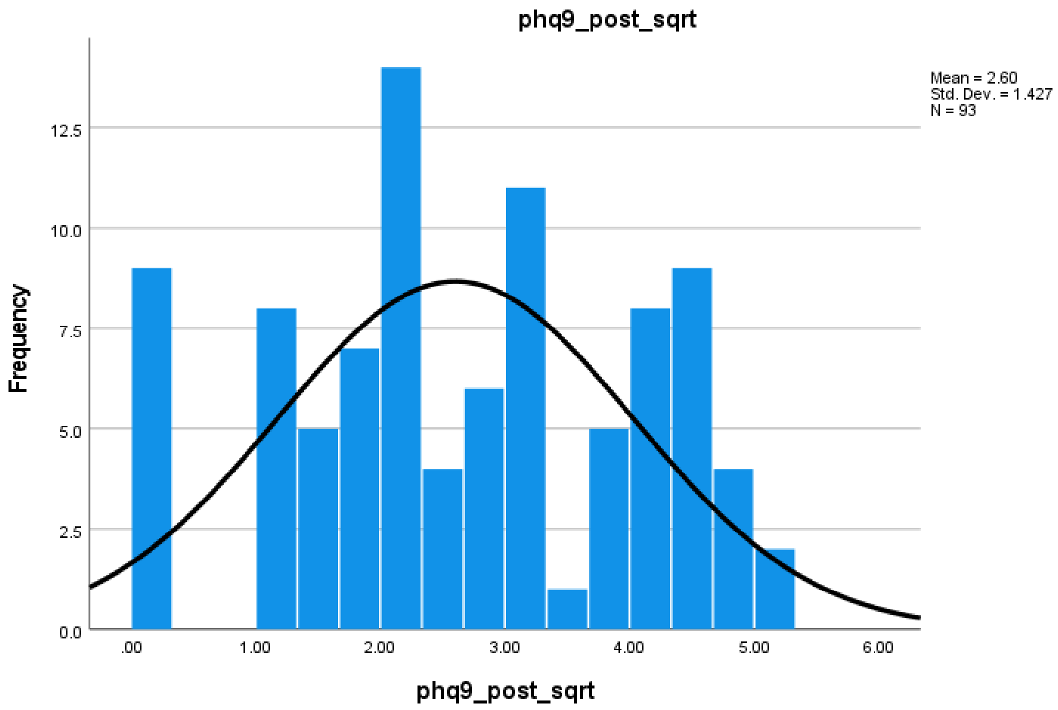


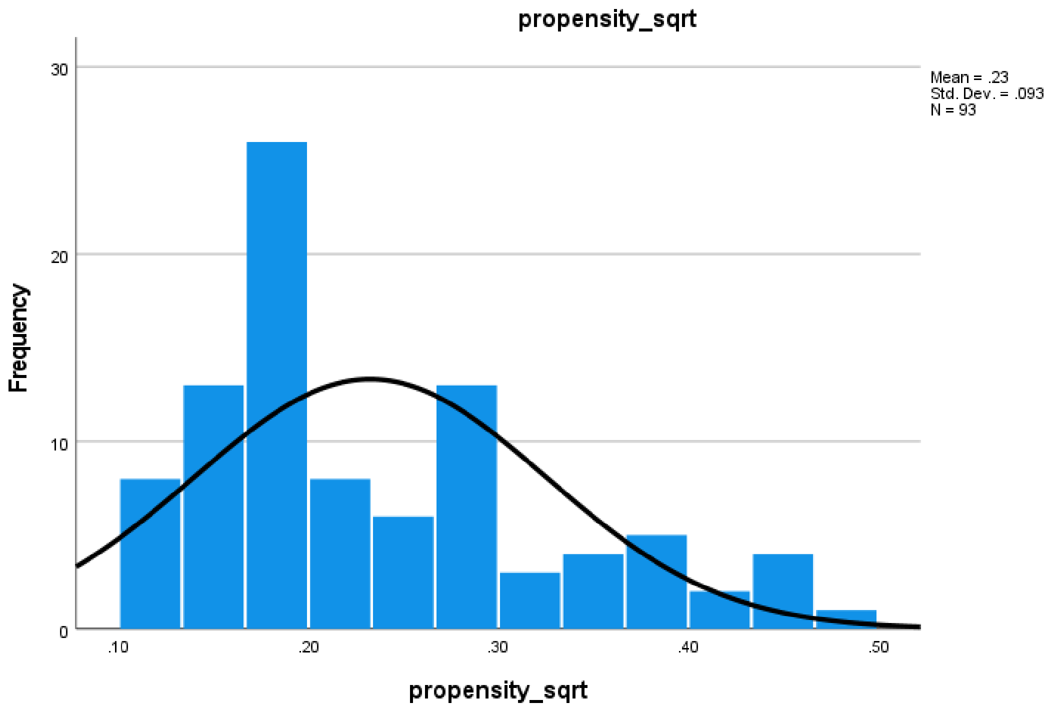

phq9_post
Mean = 8.78
Std. Dev. = 7.391
N = 93
Frequency
0
5
10
15
-10.00
.00
10.00
20.00
30.00
phq9_post


propensity_score
Mean = .06
Std. Dev. = .051
N = 93
Frequency
0
10
20
30
.00
.05
.10
.15
.20
.25
propensity_score

APPENDIX I – Figure 3

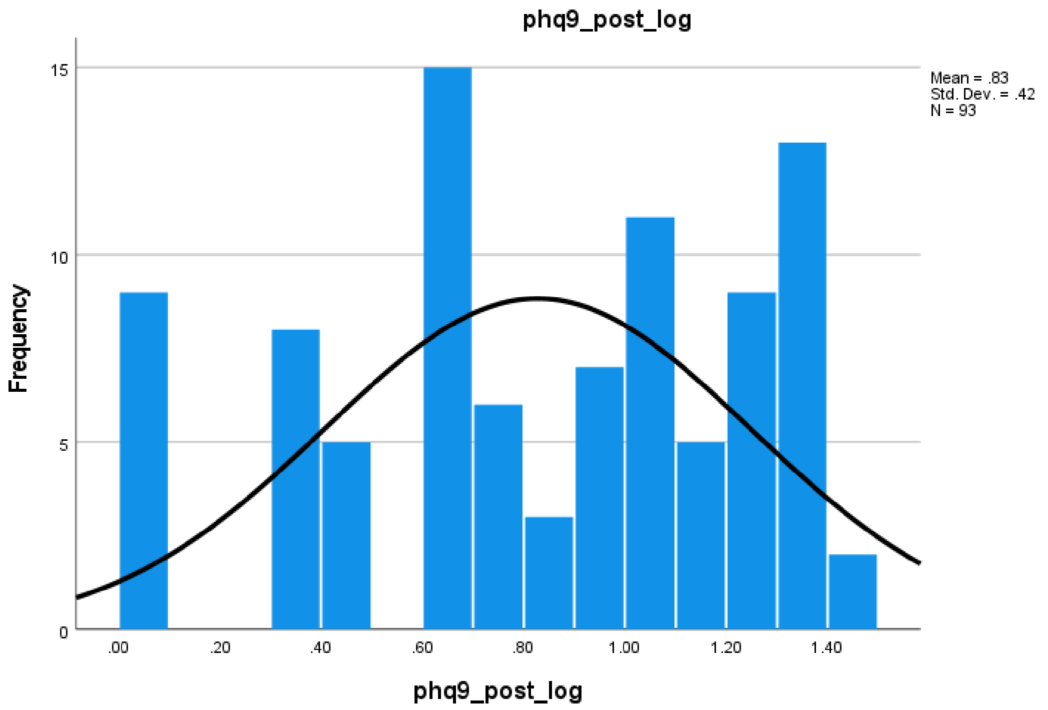
Histograms Displaying the Distributions of the Square Root Transformed Variables

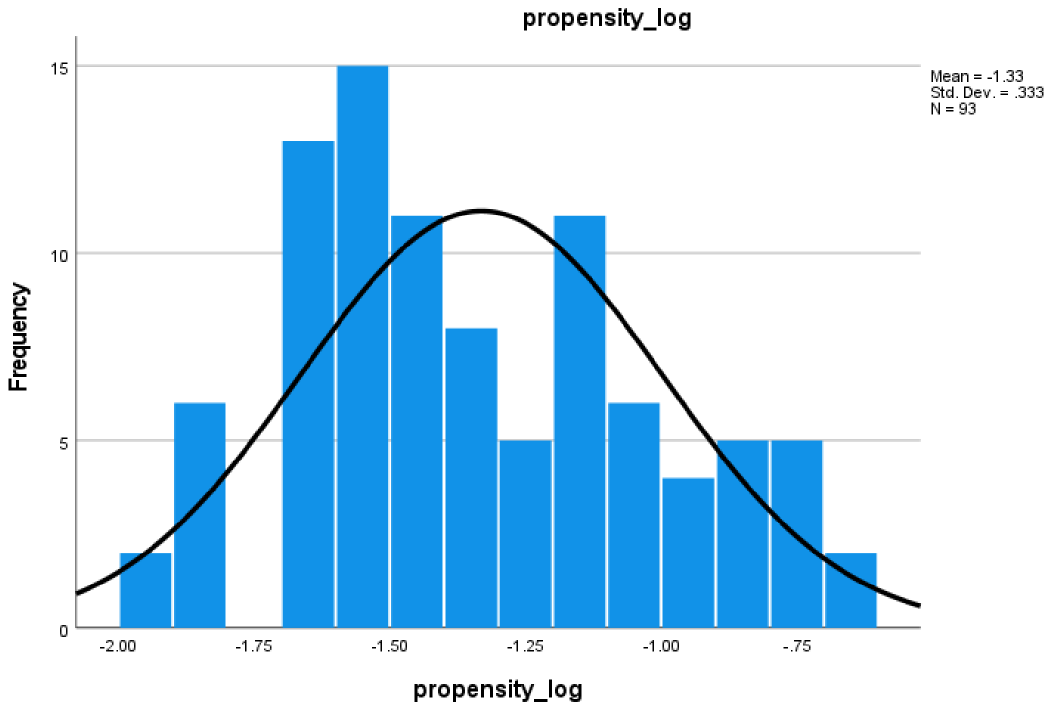




APPENDIX I – Figure 4

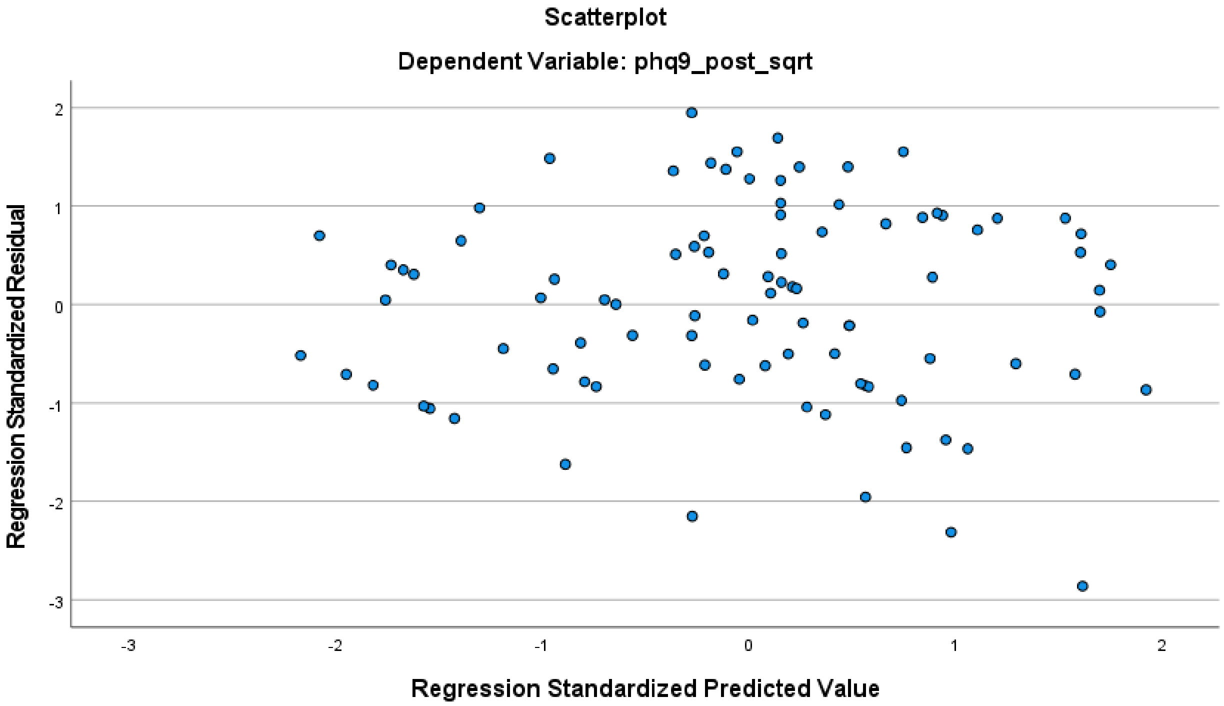
Histograms Displaying the Distributions of the Logarithm Transformed Variables





APPENDIX I – Figure 5

Scatterplot to Depict the Standardised Residuals Plotted Against Standardised Predicted Post-Treatment PHQ-9 Scores Square Root Transformed, with Propensity Score Logarithm Transformed



APPENDIX J

Missing Data in Chapter 4 Dataset

APPENDIX J - Table 1

Proportion of Values Missing for Each Variable in the Training and Validation Samples

Variable	Training sample (*N* = 855)	Validation sample (*N* = 464)
Age	0.00%	0.00%
Disability	78.36%	89.87%
Employment (pre-treatment)	7.95%	0.65%
Employment (post-treatment)	7.72%	11.85%
Ethnicity	4.91%	4.09%
GAD-7 (pre-treatment)	0.00%	0.00%
GAD-7 (post-treatment)	0.23%	0.65%
GAD7 (first Tf-CBT session)	4.44%	0.86%
GAD7 (last Tf-CBT session)	4.91%	3.02%
Gender	0.23%	0.00%
IESR (pre-treatment)	0.00%	0.00%
IESR (post-treatment)	50.53%	43.53%
IMD Decile	4.56%	4.53%
Long Term Condition	33.10%	12.50%
Medication (pre-treatment)	12.63%	3.66%
Medication (post-treatment)	12.40%	7.54%
PHQ9 (pre-treatment)	0.00%	0.00%
PHQ9 (post-treatment)	0.23%	0.65%
PHQ9 (first Tf-CBT session)	4.21%	0.86%
PHQ9 (last Tf-CBT session)	4.56%	3.02%
N Tf-CBT sessions	0.00%	0.00%
Unemployed (pre-treatment)	7.95%	0.65%
Unemployed (post-treatment)	7.72%	11.85%
WSAS (pre-treatment)	4.80%	0.22%
WSAS (post-treatment)	7.60%	1.29%
WSAS (first Tf-CBT session)	13.57%	2.16%
WSAS (last Tf-CBT session)	16.84%	3.66%

Note. GAD-7 = Generalised Anxiety Disorder 7; IES-R = Impact of Event Scale – Revised; IMD = Index of Multiple Deprivation; PHQ-9 = Patient Health Questionnaire 9; Tf-CBT = Trauma-focussed Cognitive Behavioural Therapy; WSAS = Work and Social Adjustment Scale.
All variables in except for Disability were entered into multiple imputation models for the training and validation samples.

APPENDIX K

Correlation Matrices for Chapter 4 Dataset

Supplementary Table 4.2

Spearman's Correlations Between Variables in the Training Sample Before Imputation

	Age	Ethnicity	GAD-7 (first Tf-CBT session)	GAD-7 (last Tf-CBT session)	GAD-7 (post)	GAD-7 (pre)	Gender	IES-R (post)	IES-R (pre)	IMD decile	LTC	Medication (post)	Medication (pre)	*N* Tf-CBT sessions	PHQ-9 (first Tf-CBT session)	PHQ-9 (last Tf-CBT session)	PHQ-9 (post)	PHQ-9 (pre)	Unemployed (post)	Unemployed (pre)	WSAS (first Tf-CBT session)	WSAS (last Tf-CBT session)	WSAS (post)	WSAS (pre)
Age	-	-.08	.02	-.07	-.07	.04	-.12	.00	.07	.10	.29	.07	.11	.17	.06	-.05	-.05	.06	.05	.02	.07	-.01	.02	.04
Ethnicity	-.08	-	.01	.00	.00	.02	-.03	.09	.02	-.18	.01	-.05	-.11	-.08	.05	.01	.02	.02	.07	.09	.04	.05	.04	.04
GAD-7 (first Tf-CBT session)	.02	.01	-	.40	.40	.61	-.03	.24	.48	-.10	.07	.21	.23	-.01	.72	.38	.37	.50	.25	.25	.46	.32	.31	.37
GAD-7 (last Tf-CBT session)	-.07	.00	.40	-	.99	.34	-.07	.71	.42	-.16	.07	.30	.20	-.23	.41	.91	.90	.33	.34	.30	.36	.78	.74	.27
GAD-7 (post)	-.07	.00	.40	.99	-	.34	-.08	.72	.43	-.16	.06	.31	.19	-.24	.41	.90	.91	.34	.33	.31	.38	.77	.75	.28
GAD-7 (pre)	.04	.02	.61	.34	.34	-	-.04	.21	.42	-.11	.04	.19	.22	.00	.54	.33	.32	.67	.17	.18	.38	.28	.29	.48
Gender	-.12	-.03	-.03	-.07	-.08	-.04	-	-.07	-.05	.00	-.04	-.08	-.08	.05	-.15	-.10	-.12	-.14	-.13	-.12	-.13	-.09	-.12	-.12
IES-R (post)	.00	.09	.24	.71	.72	.21	-.07	-	.42	-.09	.17	.31	.14	-.06	.31	.71	.73	.26	.27	.29	.28	.67	.67	.17
IES-R (pre)	.07	.02	.48	.42	.43	.42	-.05	.42	-	-.10	.08	.24	.20	-.04	.46	.41	.42	.38	.28	.26	.43	.40	.42	.32
IMD decile	.10	-.18	-.10	-.16	-.16	-.11	.00	-.09	-.10	-	-.04	-.08	-.06	.12	-.14	-.16	-.17	-.17	-.29	-.31	-.18	-.20	-.18	-.18
LTC	.29	.01	.07	.07	.06	.04	-.04	.17	.08	-.04	-	.12	.16	.05	.12	.07	.07	.09	.17	.12	.11	.09	.10	.12
Medication (post)	.07	-.05	.21	.30	.31	.19	-.08	.31	.24	-.08	.12	-	.54	.01	.31	.35	.36	.32	.25	.23	.25	.34	.33	.24
Medication (pre)	.11	-.11	.23	.20	.19	.22	-.08	.14	.20	-.06	.16	.54	-	.03	.29	.23	.23	.30	.21	.19	.24	.24	.24	.26
N Tf-CBT sessions	.17	-.08	-.01	-.23	-.24	.00	.05	-.06	-.04	.12	.05	.01	.03	-	.00	-.25	-.26	-.03	-.08	-.07	.00	-.22	-.20	-.03
PHQ-9 (first Tf-CBT session)	.06	.05	.72	.41	.41	.54	-.15	.31	.46	-.14	.12	.31	.29	.00	-	.48	.48	.68	.31	.31	.59	.41	.40	.48
PHQ-9 (last Tf-CBT session)	-.05	.01	.38	.91	.90	.33	-.10	.71	.41	-.16	.07	.35	.23	-.25	.48	-	.98	.40	.36	.33	.39	.81	.77	.31
PHQ-9 (post)	-.05	.02	.37	.90	.91	.32	-.12	.73	.42	-.17	.07	.36	.23	-.26	.48	.98	-	.40	.37	.34	.40	.80	.79	.32
PHQ-9 (pre)	.06	.02	.50	.33	.34	.67	-.14	.26	.38	-.17	.09	.32	.30	-.03	.68	.40	.40	-	.27	.27	.51	.38	.38	.60
Unemployed (post)	.05	.07	.25	.34	.33	.17	-.13	.27	.28	-.29	.17	.25	.21	-.08	.31	.36	.37	.27	-	.72	.31	.41	.39	.26
Unemployed (pre)	.02	.09	.25	.30	.31	.18	-.12	.29	.26	-.31	.12	.23	.19	-.07	.31	.33	.34	.27	.72	-	.31	.34	.35	.24

WSAS (first Tf-CBT session)	.07	.04	.46	.36	.38	.38	-.13	.28	.43	-.18	.11	.25	.24	.00	.59	.39	.40	.51	.31	.31	-	.55	.56	.65
WSAS (last Tf-CBT session)	-.01	.05	.32	.78	.77	.28	-.09	.67	.40	-.20	.09	.34	.24	-.22	.41	.81	.80	.38	.41	.34	.55	-	.99	.41
WSAS (post)	.02	.04	.31	.74	.75	.29	-.12	.67	.42	-.18	.10	.33	.24	-.20	.40	.77	.79	.38	.39	.35	.56	.99	-	.41
WSAS (pre)	.04	.04	.37	.27	.28	.48	-.12	.17	.32	-.18	.12	.24	.26	-.03	.48	.31	.32	.60	.26	.24	.65	.41	.41	-

Supplementary Table 4.3

Spearman's Correlations Between Variables in the Validation Sample Before Imputation

	Age	Ethnicity	GAD-7 (first Tf-CBT session)	GAD-7 (last Tf-CBT session)	GAD-7 (post)	GAD-7 (pre)	Gender	IES-R (post)	IES-R (pre)	IMD decile	LTC	Medication (post)	Medication (pre)	*N* Tf-CBT sessions	PHQ-9 (first Tf-CBT session)	PHQ-9 (last Tf-CBT session)	PHQ-9 (post)	PHQ-9 (pre)	Unemployed (post)	Unemployed (pre)	WSAS (first Tf-CBT session)	WSAS (last Tf-CBT session)	WSAS (post)	WSAS (pre)
Age	-	-.06	-.01	-.04	-.02	-.03	-.10	-.06	-.02	.13	.24	.14	.03	.10	.02	-.05	-.03	.01	.01	-.02	.07	-.03	-.03	-.01
Ethnicity	-.06	-	.04	.05	.05	.06	-.12	.13	.19	-.28	-.13	.06	.09	.01	.09	.08	.08	.13	.06	.11	.06	.07	.08	.15
GAD-7 (first Tf-CBT session)	-.01	.04	-	.45	.45	.62	.06	.34	.40	-.04	.02	.13	.16	-.02	.70	.39	.39	.50	.20	.19	.46	.38	.37	.35
GAD-7 (last Tf-CBT session)	-.04	.05	.45	-	.99	.29	.02	.79	.30	-.11	-.01	.08	.08	-.33	.44	.89	.88	.35	.33	.31	.32	.79	.79	.27
GAD-7 (post)	-.02	.05	.45	.99	-	.29	.01	.79	.30	-.11	.00	.08	.08	-.34	.44	.89	.90	.36	.34	.31	.33	.80	.79	.27
GAD-7 (pre)	-.03	.06	.62	.29	.29	-	.02	.27	.41	-.09	.05	.14	.17	.00	.51	.26	.26	.64	.25	.24	.33	.24	.24	.45
Gender	-.10	-.12	.06	.02	.01	.02	-	-.04	.04	-.05	-.07	-.01	-.02	.07	-.02	-.03	-.03	-.04	-.09	-.07	-.04	-.06	-.06	-.09
IES-R (post)	-.06	.13	.34	.79	.79	.27	-.04	-	.35	-.21	-.03	.12	.04	-.13	.41	.78	.78	.34	.36	.35	.27	.72	.72	.32
IES-R (pre)	-.02	.19	.40	.30	.30	.41	.04	.35	-	-.12	-.01	.18	.18	-.03	.43	.32	.32	.45	.28	.25	.29	.32	.31	.32
IMD decile	.13	-.28	-.04	-.11	-.11	-.09	-.05	-.21	-.12	-	.11	-.10	-.12	.04	-.02	-.09	-.10	-.10	-.11	-.15	-.02	-.08	-.07	-.13
LTC	.24	-.13	.02	-.01	.00	.05	-.07	-.03	-.01	.11	-	.01	-.02	.00	.07	.02	.03	.04	.09	.04	.01	-.01	.00	-.02
Medication (post)	.14	.06	.13	.08	.08	.14	-.01	.12	.18	-.10	.01	-	.46	.00	.15	.10	.10	.13	.24	.13	.12	.09	.09	.14
Medication (pre)	.03	.09	.16	.08	.08	.17	-.02	.04	.18	-.12	-.02	.46	-	-.01	.18	.08	.07	.22	.22	.15	.18	.09	.08	.23
N Tf-CBT sessions	.10	.01	-.02	-.33	-.34	.00	.07	-.13	-.03	.04	.00	.00	-.01	-	.00	-.35	-.34	.00	-.14	-.15	.04	-.25	-.23	.05
PHQ-9 (first Tf-CBT session)	.02	.09	.70	.44	.44	.51	-.02	.41	.43	-.02	.07	.15	.18	.00	-	.51	.50	.72	.35	.32	.59	.45	.45	.49
PHQ-9 (last Tf-CBT session)	-.05	.08	.39	.89	.89	.26	-.03	.78	.32	-.09	.02	.10	.08	-.35	.51	-	.99	.41	.36	.36	.36	.84	.83	.30
PHQ-9 (post)	-.03	.08	.39	.88	.90	.26	-.03	.78	.32	-.10	.03	.10	.07	-.34	.50	.99	-	.42	.37	.36	.35	.83	.83	.29

PHQ-9 (pre)	.01	.13	.50	.35	.36	.64	-.04	.34	.45	-.10	.04	.13	.22	.00	.72	.41	.42	-	.34	.35	.47	.39	.38	.60
Unemployed (post)	.01	.06	.20	.33	.34	.25	-.09	.36	.28	-.11	.09	.24	.22	-.14	.35	.36	.37	.34	-	.61	.24	.33	.34	.26
Unemployed (pre)	-.02	.11	.19	.31	.31	.24	-.07	.35	.25	-.15	.04	.13	.15	-.15	.32	.36	.36	.35	.61	-	.19	.31	.30	.27
WSAS (first Tf-CBT session)	.07	.06	.46	.32	.33	.33	-.04	.27	.29	-.02	.01	.12	.18	.04	.59	.36	.35	.47	.24	.19	-	.46	.45	.62
WSAS (last Tf-CBT session)	-.03	.07	.38	.79	.80	.24	-.06	.72	.32	-.08	-.01	.09	.09	-.25	.45	.84	.83	.39	.33	.31	.46	-	.99	.41
WSAS (post)	-.03	.08	.37	.79	.79	.24	-.06	.72	.31	-.07	.00	.09	.08	-.23	.45	.83	.83	.38	.34	.30	.45	.99	-	.39
WSAS (pre)	-.01	.15	.35	.27	.27	.45	-.09	.32	.32	-.13	-.02	.14	.23	.05	.49	.30	.29	.60	.26	.27	.62	.41	.39	-

APPENDIX L

Distributions of Numeric Variables in Chapter 4 Dataset

APPENDIX L - Table 1

Descriptive Statistics for Continuous Variables in the Training Sample Following Imputation

Variable	Mean	Median	Mode	Min	Max	Skew	SE skew	Kurtosis	SE kurt'
age	38.66	37	27	16	76	0.293*	0.084	-0.787*	0.167
iesr_first	60.09	63	71	0	88	-0.839*	0.084	0.502*	0.167
iesr_last	37.90	37	22	0	88	0.084	0.084	-1.191*	0.167
phq9_first	17.45	18	21	0	27	-0.591*	0.084	-0.316	0.167
phq9_last	11.21	10	0	0	27	0.303*	0.084	-1.128*	0.167
gad7_first	15.62	17	21	0	21	-0.876*	0.084	0.103	0.167
gad7_last	10.11	9	21	0	21	0.154	0.084	-1.319*	0.167
wsas_first	22.17	23	32	0	40	-0.192*	0.084	-0.783*	0.167
wsas_last	16.09	15	0	0	40	0.321*	0.084	-1.000*	0.167

Note. _first = pre-treatment score; _last = post-treatment score; gad7 = Generalised Anxiety Disorder 7; iesr = Impact of Event Scale – Revised; phq9 = Patient Health Questionnaire 9; wsas = Work and Social Adjustment Scale.

* An absolute skew or kurtosis value greater than twice its standard error (SE) indicates significant skew or kurtosis

APPENDIX L - Table 2

Descriptive Statistics for Continuous Variables in the Validation Sample Following Imputation

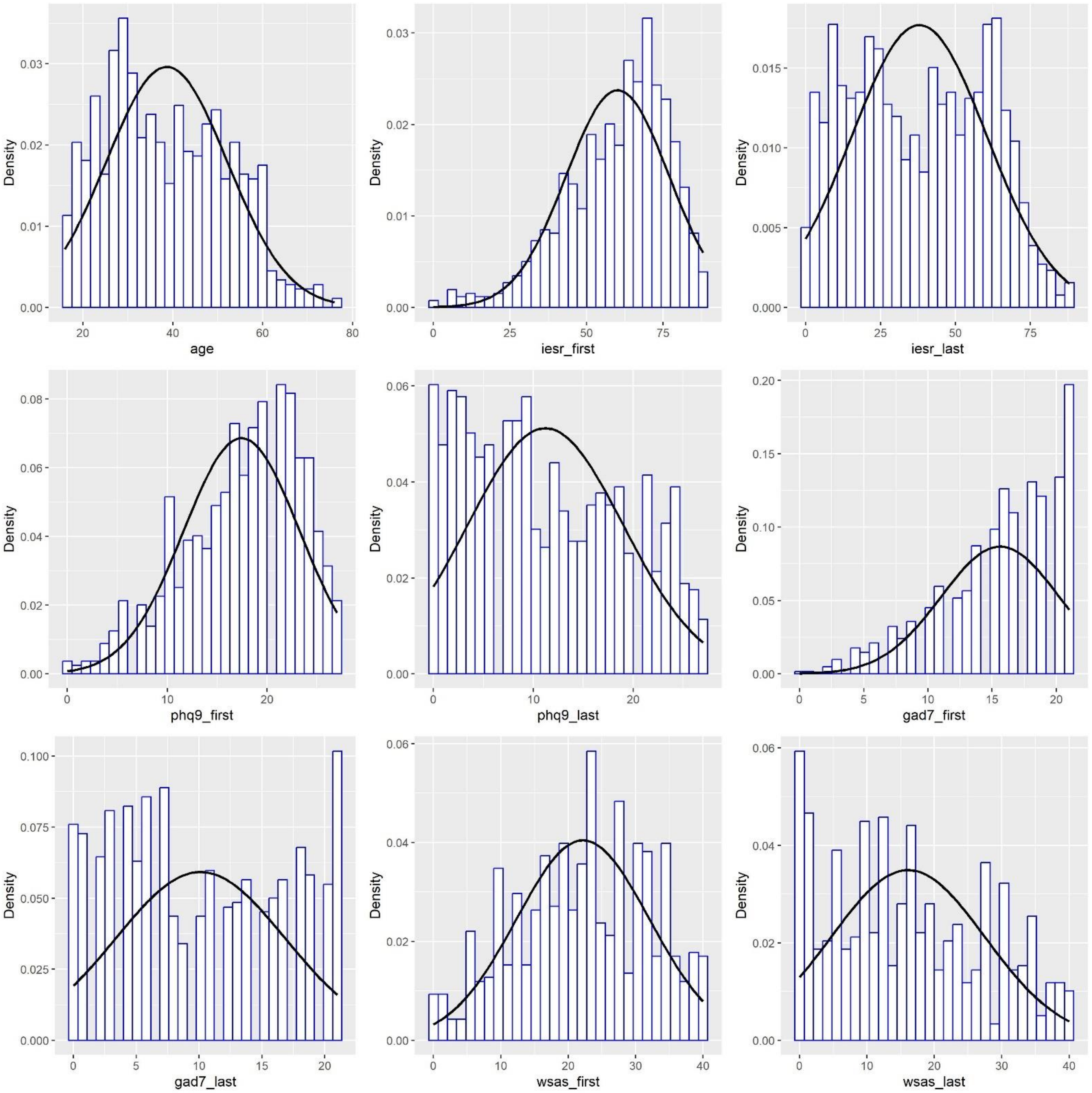
Variable	Mean	Median	Mode	Min	Max	Skew	SE skew	Kurtosis	SE kurt'
age	38.91	37.5	35	17	73	0.335*	0.113	-0.779*	0.226
iesr_first	60.15	64	70	0	88	-1.077*	0.113	1.142*	0.226
iesr_last	37.80	33	12	0	87	0.216	0.113	-1.251*	0.226
phq9_first	17.14	18	18	0	27	-0.513*	0.113	-0.286	0.226
phq9_last	10.56	9	8	0	27	0.439*	0.113	-0.939*	0.226
gad7_first	15.74	16	21	3	21	-0.805*	0.113	0.022	0.226
gad7_last	9.75	8.5	7	0	21	0.279*	0.113	-1.187*	0.226
wsas_first	24.19	26	28	0	40	-0.427*	0.113	-0.659*	0.226
wsas_last	16.05	15	0	0	40	0.349*	0.113	-1.114*	0.226

Note. _first = pre-treatment score; _last = post-treatment score; gad7 = Generalised Anxiety Disorder 7; iesr = Impact of Event Scale – Revised; phq9 = Patient Health Questionnaire 9; wsas = Work and Social Adjustment Scale.

* An absolute skew or kurtosis value greater than twice its standard error (SE) indicates significant skew or kurtosis

APPENDIX L - Figure 1

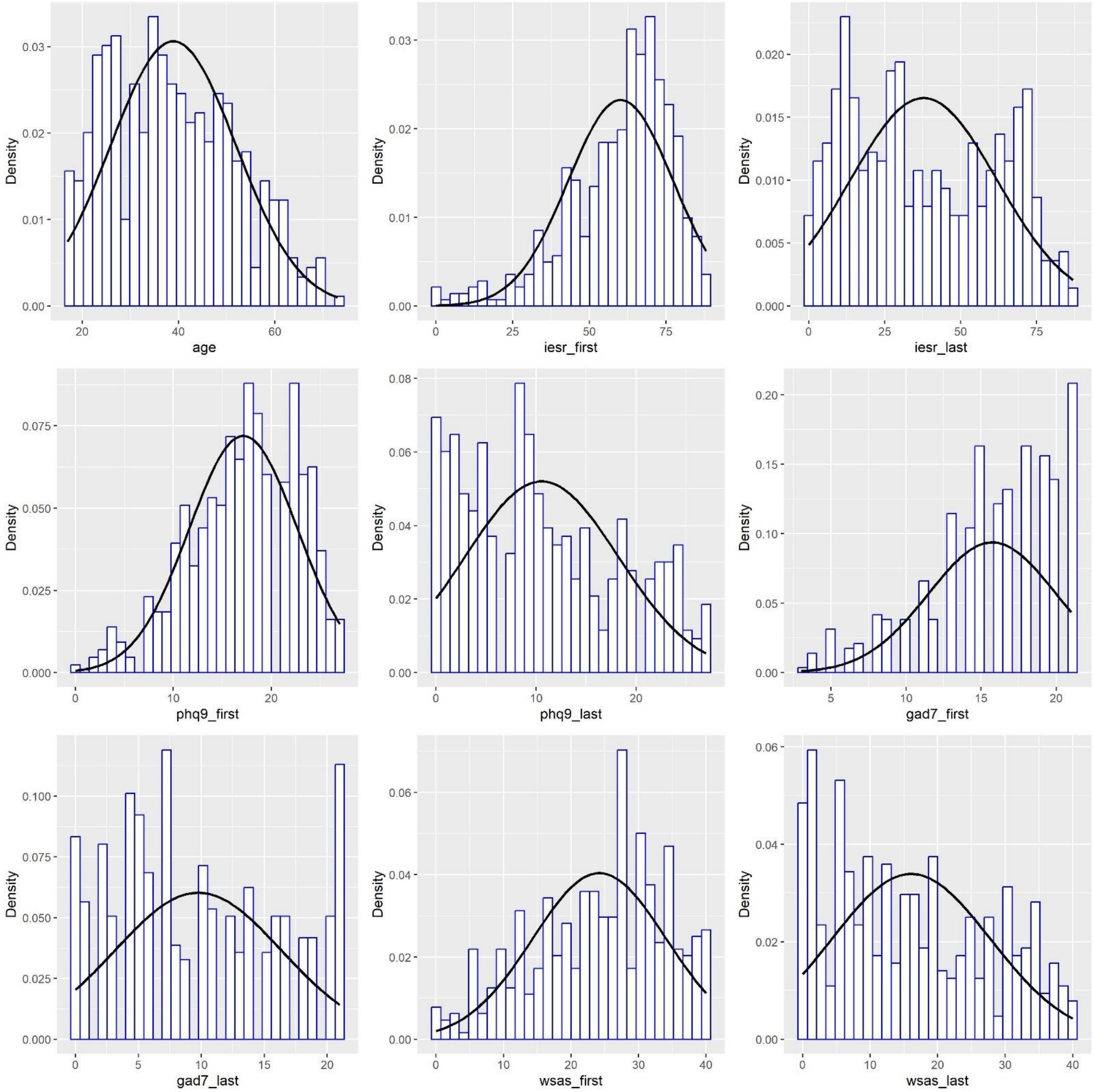
Histograms Displaying the Density and Normal Probability Curve for Continuous Variables in the Training Sample



Note. _first = pre-treatment score; _last = post-treatment score; gad7 = Generalised Anxiety Disorder 7; iesr = Impact of Event Scale – Revised; phq9 = Patient Health Questionnaire 9; wsas = Work and Social Adjustment Scale.

APPENDIX L - Figure 2

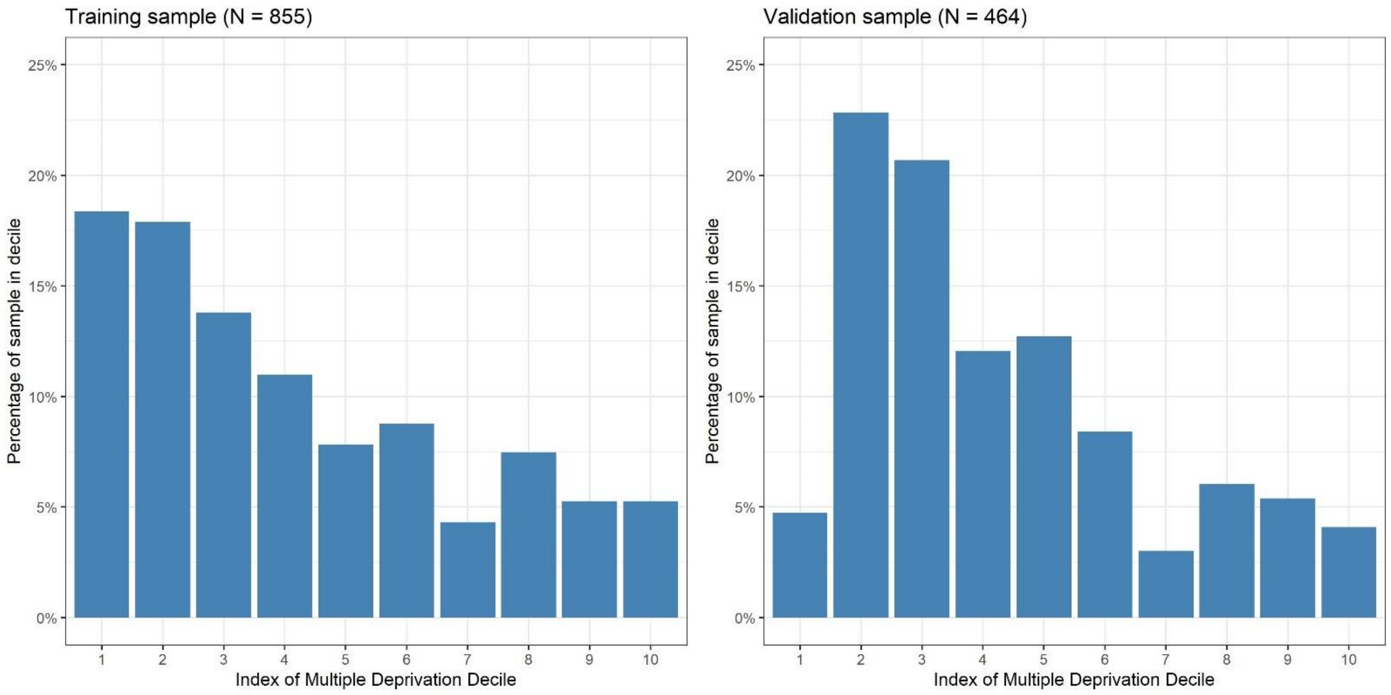
Histograms Displaying the Density and Normal Probability Curve for Continuous Variables in the Validation Sample



Note. _first = pre-treatment score; _last = post-treatment score; gad7 = Generalised Anxiety Disorder 7; iesr = Impact of Event Scale – Revised; phq9 = Patient Health Questionnaire 9; wsas = Work and Social Adjustment Scale.

APPENDIX L – Figure 3

Bar Charts Displaying the Percentage of Participants in Each Index of Multiple Deprivation Decile in the Training and Validation Sample



Training sample (N = 855)

Validation sample (N = 464)

APPENDIX M

Prediction Methods Applied in Chapter 4

APPENDIX M – Table 1

Prediction Methods Tested in Chapter 4

Method	Family	R packages	Details
Linear Regression (LR)	Non-ML comparator	caret, method = lm	A relatively simple, traditional statistical method that models the association between a set of predictor variables and an outcome variable. LR estimates a coefficient for each predictor by minimising the sum of squared errors between predicted and observed scores. The linear combination of coefficients can then be applied to predict outcomes in new data (Su et al., 2012).
Elastic Net (EN)	Penalised regressions	caret, method = glmnet	Penalised regression methods apply shrinkage penalties to improve generalisability and interpretability of regression models (Tibshirani, 1996; Zou & Hastie, 2005). EN combines the regularisation penalties of Ridge regression, which shrinks all coefficients towards zero to prevent overfitting, and the Least Absolute Shrinkage and Selection Operator (LASSO), which performs predictor selection by shrinking the coefficient to 0 for variables with little predictive value or high multicollinearity (correlation with other predictors). The hyperparameter α controls the blend between Ridge and LASSO penalisation, where α = 0 is pure Ridge regression and α = 1 is pure LASSO.
Random Forest (RF)	Decision trees	caret, method = rf	Decision trees sequentially divide the data at the most informative threshold on important predictor variables, to produce a simple and interpretable model (resembling a family tree) that can implicitly handle non-linear relationships, complex interactions, and predictor selection. To reduce overfitting, RF uses bootstrap aggregating, or *bagging*, to train a large ensemble of decision trees on subsamples of the training data, and then averages across the

			ensemble to arrive at a prediction (Breiman, 2001a). The disadvantage is that the RF model is more difficult to interpret, as it is composed of hundreds or thousands of different decision tree models that each contribute to the prediction.
Boosted Generalised Linear Model (BoostGLM)	Boosted models	caret, method = glmboost	Boosted models combine many "weak" (in this case linear) models referred to as *base-learners*, to form a "strong" ensemble model. However, instead of training base-learners in parallel and taking an average of model outcomes (as in bagging methods such as random forest), boosted models sequentially update the prediction model, with each base-learner trying to reduce the prediction error of the preceding base-learner(s). Each base-learner only includes a subset of the model predictors, and the hyperparameter mstop stops the algorithm before overfitting the training data, implicitly performing predictor selection and regularisation (Hofner et al., 2014).
Bayesian Generalised Linear Model (BayesGLM)	Bayesian models	caret, method = bayesglm	Bayesian models apply Bayesian probability theory to prediction problems. Unlike frequentist statistics, Bayesian methods begin by proposing a mathematical model (or *prior distribution*) based on prior knowledge (e.g., past findings or expert opinion), before analysing the data. The prior distribution is then updated in response to the data analysis to give a new model referred to as the *posterior distribution*. In this way the model incorporates prior beliefs about the relationships between variables with the information provided by the current dataset (Kruschke, 2011). In Bayesian machine learning, default priors are often supplied by the algorithm, and in the case of the bayesglm method (Gelman et al., 2008), weak, minimally informative priors are used to regularise the model coefficients (i.e., shrink them towards zero to limit overfitting).
Radial Support Vector Machine (RSVM)	Linear models	caret, method = svmRadial	RSVM is a support vector machine (Vapnik et al., 1995) with a radial basis function kernel. With a continuous outcome the support vector machine algorithm performs support vector regression. Each observation in the dataset exists in a space with as many dimensions as there are variables, and support vector regression seeks to fit the flattest possible hyperplane

			that minimises the error between the hyperplane and each observation. The flatter the hyperplane the more generalisable the model is likely to be, so to increase generalisability the errors that fall within a particular margin around the hyperplane are ignored. The width of the error margin is controlled by the hyperparameter C. The radial basis function kernel transforms the data in such a way that allows the support vector machine to fit complex, non-linear relationships (Hastie et al., 2009; Smola & Schölkopf, 2004).
Multi-Layer Perceptron (MLP)	Deep learning	caret, method = monmlp	Monotone MLP (monmlp) is a neural network from the *deep learning* family, which fits one- and two-layer perceptron neural networks with an optional monotone constraint (H. Zhang & Zhang, 1999). A perceptron is a construct inspired by neurons in the human brain, in which a set of inputs pass through a mathematical function to produce a set of outputs. Perceptron are organised in interconnected layers, and the inputs (e.g., pre-treatment data) may pass through multiple *hidden* layers before the output (e.g., treatment outcome prediction) is produced; hence the term *multi-layer* perceptron. Through a process called *back-propagation* the multi-layer perceptron iteratively evaluates the accuracy of the outputs and updates the functions in the hidden layers. Bootstrapping is recommended to prevent this process from overfitting. For a thorough introduction to neural networks see Hastie et al. (2009). The optional monotone constraint allows the user to specify whether outputs of specific variables can only ever decrease or increase relative to another variable (e.g., constrain predicted post-treatment PTSD score so that it cannot decrease as pre-treatment PTSD score increases). monmlp was chosen due to the findings of Giesemann et al. (2023), but for fairness of comparison between models in the present study, default settings were applied. Therefore, the monmlp model trained in the present study was an MLP without monotone constraints.
Bayesian Regularised Neural Network (BRNN)	Deep learning/	caret, method = brnn	Like MLP (above), BRNN is a multi-layer neural network. However, instead of using back-propagation, BRNN uses priors to regularise the model coefficients (see BayesGLM), as this

	Bayesian models		may produce a more generalisable model (Foresee & Hagan, 1997). As such, BRNN belongs to both the deep learning family and the Bayesian family.
Genetic Regression (GR)	Evolutionary algorithms	glmulti, method = g caret, method = lm	GR uses a genetic algorithm to select predictors for a regression model. Genetic algorithms belong to the *evolutionary algorithms* family, which employ processes inspired by the theory of evolution to “evolve” the best solution to a problem over many “generations” (Mitchell, 1998). In contrast to the other methods tested in this study, there was a preliminary stage to training the genetic regression model: Important predictors were identified using the genetic algorithm option (g) in the glmulti package (with confsetsize = 512 and default settings for all other parameters). Predictors with importance > .8 were then entered into a linear regression model (lm), which was trained and internally cross-validated using the caret package in the same way as the other models. This was done to replicate the methods used by Deisenhofer et al. (2018) when developing the model tested in Chapter 3, while retaining the same cross-validation procedure for comparability with other models in this study.