Oparina, Ekaterina; Krekel, Christian; Srisuma, Sorawoot

Working Paper

Talking Therapy: Impacts of a Nationwide Mental Health Service in England

IZA Discussion Papers, No. 16839

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Oparina, Ekaterina; Krekel, Christian; Srisuma, Sorawoot (2024) : Talking Therapy: Impacts of a Nationwide Mental Health Service in England, IZA Discussion Papers, No. 16839, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/295862

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16839

Talking Therapy: Impacts of a Nationwide Mental Health Service in England

Ekaterina Oparina ®
Christian Krekel ®
Sorawoot Srisuma ®

MARCH 2024

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16839

Talking Therapy: Impacts of a Nationwide Mental Health Service in England

Ekaterina Oparina ®
CEP and LSE

Christian Krekel ®
CEP, LSE and IZA

Sorawoot Srisuma ®
National University of Singapore and University of Surrey

MARCH 2024

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Talking Therapy: Impacts of a Nationwide Mental Health Service in England*

Common mental health problems impose significant costs on individuals and societies, yet healthcare systems often overlook them. We provide the first causal evidence on the effectiveness of a pioneering, nationwide mental health service for treating depression and anxiety disorders in England using non-experimental data and methods. We exploit variations in waiting times to identify treatment effects, based on a novel dataset of over one million patients that well represent the English population. Our findings show that treatment improved mental health and reduced impairment in work and social life. We also provide suggestive evidence of enhanced employment. However, effects vary across patients, services, and areas. The programme is cost-effective and provides a blueprint for treating mental health in other countries.

JEL Classification: C31, C32, D61, I12, I38

Keywords: policy evaluation, mental health, psychological therapies, quasi-natural experiment, machine learning, cost-benefit analysis

Corresponding author:
Ekaterina Oparina
London School of Economics (LSE)
Houghton St
London WC2A 2AE
United Kingdom
E-mail: e.oparina@lse.ac.uk

* The symbol ® indicates that the author order was determined by the American Economic Association's Author Randomization Tool. We are grateful to Richard Layard and David Clark for valuable discussions, comments, and suggestions throughout this project. Niall Maher and Isaac Parkes provided excellent research assistance. We thank Martin Knapp, Henry Overman, and conference participants at the Annual Conference of the American Economic Association as well as seminar participants at the LSE Wellbeing Seminar and the Mental Health and Economic Status Conference (University of Warwick) for comments. This work was supported by the ESRC [Grant Number: ES/W002094/1].

1 Introduction

With one in four adults experiencing at least one diagnosable mental health problem in any given year, mental ill health is the largest single cause of disability and illness in the UK, accounting for an estimated 91 million working days per year lost, nearly 40% of people on disability benefits, and about one third of GPs' time (R. Layard, 2016). It is estimated that, in the UK, the economic costs of mental health problems range between 3.5% and 4% of GDP annually. In the US, they are estimated to be even higher, ranging between 6% and 10.4% (Arias, Saxena, and Verguet, 2022).

Although mental ill health imposes a substantial burden on individuals and their families, as well as on economies and societies as a whole, it is seldom prioritised in healthcare systems. Meanwhile, substantial progress has been made on the developments of evidence-based psychological therapies for a wide range of mental health problems. Indeed, the consensus amongst clinicians and practitioners is that such therapies can make a huge difference to patients and their lives (A. Roth and D. Fonagy, 2005; Lambert, 2013; Nathan and Gorman, 2015). Importantly, patients themselves report a threefold preference for therapy over medication(McHugh et al., 2013).[1]

To confront this situation, in 2008, the UK Government launched a large-scale mental health service in England to make evidence-based psychological therapies more widely available within the *National Health Service (NHS)*, its universal public healthcare system. This nationwide programme, which was then, and still is, the largest in the world, is called the *Improving Access to Psychological Therapies (IAPT)* programme.[2] To date, IAPT has trained over 10,500 new therapists and treated over seven million patients (more than 13% of the English population), primarily via cognitive behavioural therapies (CBT), in a stepped-care model with both low and high-intensity treatments (NHS, 2021).[3]

Today, IAPT is widely regarded as a success and is being replicated in other

[1] Compared to medication, such as antidepressants, therapy is used much less in the UK (McManus, Bebbington, and Jenkins, 2016) and in the US (Marcus and Olfson, 2010). For example, only 10% of adults with common mental health disorders such as depression or anxiety in the UK received therapy in 2007 (i.e. the year prior to the launch of the programme studied in this paper), and only 5% therapies that are empirically supported (R. Layard and Clark, 2014).

[2] The programme has recently been renamed *NHS Talking Therapies for Anxiety and Depression*.

[3] The psychological therapies provided by IAPT are recommended by the *National Institute for Health and Care Excellence (NICE)* in the UK and, hence, supported by extensive body of causal research on their effectiveness.

countries, e.g. Norway, Sweden, and Australia (Clark, 2019). However, until now, the causal effects of the programme on patients' treatment outcomes have never been estimated beyond small-scale RCTs. Moreover, existing results from correlational studies suggest substantial differences in outcomes between patients of different demographic and socio-economic characteristics as well as different geographical locations. Indeed, a priority for the NHS at the moment is to understand why treatment works well for some patients but not for others.[4]

In this paper, we provide the first causal evidence on the effectiveness of being treated within the IAPT programme on patients' mental health outcomes, using a unique dataset that includes data on over one million patients – all individuals who started their treatment between April 2016 and December 2018. Our quasi-experimental identification strategy, which relies institutional knowledge, supported by empirical evidence on the programme's implementation characteristics, allows us to estimate average as well as heterogeneous treatment effects for a representative sample several orders of magnitude larger than a typical mental-health focused RCT.[5]

For identification, we rely on the oversubscription of patients to the programme. The latter creates exogenous variations in waiting times, as more patients are referred to treatment than can be quickly treated, across services and over time. This enables us to identify the causal treatment effects of the IAPT programme by comparing the changes in mental health of patients who were awaiting for the start of treatment to those of patients who completed treatment during the same time period. We estimate average treatment effects using a regression framework, and heterogeneous treatment effects using nonparametric methods with and without machine learning (generalised random forests). We combine our comprehensive patient-level data with regional data on service characteristics from NHS Digital as well as socio-economic characteristics of local areas from the Office for National Statistics (ONS) in the UK.

We find that, relative to waitlisted patients in our quasi-experimental control group, treated patients' mental health is significantly more likely to have *reliably*

[4]The NHS's *Mental Health Implementation Plan 2019/20 – 2023/24* states that one of IAPT's top priorities is "[...] reducing geographic variation between services and reducing inequalities in [...] outcomes for particular population groups" (NHS England, 2019).

[5]We have benefited greatly from discussions with Richard Layard and David Clark, the founders of the IAPT programme, whose insights on the programme's key considerations and how it operates help inform our identification strategy.

improved, with a *reliably recovery* rate from mental ill health of about 43%. This is regardless of the intensity of treatment (i.e. low or high-intensity treatment), suggesting that the allocation of patients by therapists to different treatment intensities results in an appropriate patient-therapy fit. Conversely, we find that treated patients' mental health is significantly less likely to have *reliably deteriorated*. The latter is a result worth highlighting as novel empirical evidence that addresses recent concerns that well-intended psychological interventions may inadvertently cause harm (see, for example, Harvey et al. (2023) on dialectic behavioural therapy and its adverse social and emotional wellbeing outcomes amongst youth). Further, we find strong reductions in adverse mental health symptoms, as measured by a scale for depression symptoms, PHQ-9, and a scale for anxiety-disorder symptoms, GAD-7 (approximately 5.1 and 4.8 points, or 93% and 110% SD, respectively, of the pre-treatment scores in the treatment group). Moreover, there is a noteworthy reduction of 116% SD in an overall index capturing mental ill health.

Finally, there is evidence of positive, short-term ripple effects on work and social life. Amongst those who were initially unemployed or on long-term sick leave, treated patients are significantly more likely to report being employed at the end of treatment (an increase of about three percentage points) and significantly less likely to receive statutory sick pay (a decrease of about three percentage points).

However, as anticipated, there are substantial heterogeneities in the treatment effect of the programme. Overall, the categories of patients that typically have lower mental health outcomes, e.g. those who live with a disability, also benefit less from the programme. Area deprivation is negatively related to patient outcomes, whereas the funding of services is positively related. We also find that compliance with official guidelines and recommendations about the selection of therapy types is associated with better patient outcomes.

Our results are robust to different definitions of treatment and control group when varying treatment and corresponding waiting time durations, to different disease subsets when selectively including or excluding certain mental health problems, and to using a wide range of alternative models and outcomes.

The use of waitlists to identify treatment effects in economics is not new. An early contribution is found in Berger and Black (1992). This idea has also been implemented in an experimental setting (cf. Jacob and Ludwig, 2012; Jacob, Kapustin, and Ludwig, 2015; Finkelstein, Hendren, and Luttmer, 2019). More recent works, like ours, exploit naturally occurring waitlists due to oversubscription or excess

demand (Dague, DeLeire, and Leininger, 2017; Robles, Gross, and Fairlie, 2021; Dinerstein, Megalokonomou, and Yannelis, 2022; Hoe, 2023; Beam and Quimbo, 2023). Thus, our study adds to this quasi-experimental literature. Importantly, for such an identification strategy to be valid in our setting, treatment allocation in the IAPT programme has to strictly follow a first-come first-serve basis. This is indeed followed through in the case as IAPT, which aspires to ensure a fair treatment of patients.

Earlier evaluations of the nationwide implementation of the IAPT programme provided correlational evidence based on the comparison of patients' states before and after treatment, commingling the causal effect of treatment with natural recovery or deterioration, or other trends. The first empirical study by Clark, R. Layard, et al. (2009) evaluated two demonstration sites using before-after comparisons. The authors found a recovery rate of about 56%, which was largely maintained in a follow-up about ten months later.[6] Gyani et al. (2013) estimated the pre-post recovery rate to be 40.3% at the early stages of national rollout. Later in the rollout, recovery rates exceeded the original target of 50% (Clark, Canvin, et al., 2018).[7]

Another stream of evidence supporting the effectiveness of the programme comes from small-scale, short-run RCTs, testing new therapeutic approaches[8] or isolated components of the system.[9] Two recent RCTs show the effectiveness of IAPT-style interventions in other countries. A Norwegian study by Knapstad et al. (2020) involving 681 patients suffering from moderate depression or anxiety shows significant recovery rates and symptom reductions. In a follow-up study, Smith et al. (2024) find that former patients exhibit significantly higher incomes three years post-treatment, with a resulting benefit-cost ratio of about 4. A Spanish study involving 1,691 patients demonstrated that adding an IAPT-style psychological treatment in primary care was more (cost-)effective than treatment-as-usual (Cano-Vindel et al., 2022).

While RCTs are considered the gold standard to estimate causal treatment effects due to their controlled environment, they run on a relatively small scale. Particularly, for a nationwide programme like IAPT, outcomes from 1,000 patients may

[6] See also Richards and Suckling (2009), who also evaluated one of these sites.

[7] See J. Delgadillo et al. (2018) for area-level analysis.

[8] See P. Fonagy et al. (2019), Toffolutti et al. (2021), Clark, Wild, et al. (2022), Ehlers et al. (2023), or Strauss et al. (2023), for example

[9] See Richards, Enrique, et al. (2020) or Gruber et al. (2022), for example, andWakefield et al. (2020) for a meta-analysis of earlier RCTs.

not be representative of the general population. Moreover, the implementation of the programme in a controlled environment may inevitably differ from a nationwide rollout. In contrast, our results aspire to be policy-relevant on a national scale. Indeed, a novel feature of our study is the sheer scale of our sample, which includes over one million patients from all treatment sites in England. The size of our sample is especially informative for policy-makers interested in understanding heterogeneity in recovery rates amongst patients, services, and areas.[10]

Although mental ill health costs the taxpayer billions of dollars every year, the literature in economics has, so far, looked at mental health mostly as a by-product, for example of interventions aimed at making people move towards higher living standards (Stillman, McKenzie, and Gibson, 2009; Fryer Jr. and Katz, 2013; Ludwig et al., 2013) or of policy changes in the areas of labour, health, and social protection (Lang, 2013; Barnay and Juin, 2016; Avendano, Coulon, and Nafilyan, 2020; Ortega, 2022; Chuard, 2023). Only recently have scholars started looking at interventions and policies aimed at *directly* improving mental health amongst the general population, for example via therapy. Our work complements the current and fast-growing literature in economics that documents positive impacts of psychological therapies on various health and human capital outcomes.[11] Most of these studies find medium to strong impacts that are often lasting.[12] The evidence comes mostly from developing countries (a notable exception is Blattman, Jamison, and Sheridan (2017), who study the impact of CBT on criminal arrests in Chicago) and relies exclusively on RCTs, often with relatively small samples. The methodological difference between these papers and ours is that we take a quasi-experimental approach, which can be useful for guiding counterfactual questions on scaling up smaller pilots to the policy level (cf. List, 2022).

[10] For earlier evidence on heterogeneities in mental health outcomes, see Gyani et al. (2013), Saunders, Cape, et al. (2016) and Saunders, Buckman, and Pilling (2020), Clark, Canvin, et al. (2018), J. Delgadillo et al. (2018), and Moller et al. (2019).

[11] Examples include perinatal depression and subsequent female empowerment and investments into children's cognitive and socio-emotional skills (Baranov et al., 2020; Sevim et al., 2023b; Sevim et al., 2023a); mental health of individuals living in poor households (Barker et al., 2022); antisocial and criminal behaviour amongst economically disadvantaged youth (Blattman, Jamison, and Sheridan, 2017; Heller et al., 2017); self-image (Ghosal et al., 2022); and overall psychological and economic wellbeing (Bossuroy et al., 2022; Haushofer, Mudida, and Shapiro, 2022). Angelucci and Bennett, 2023 look at antidepressants and livelihoods support, individually and jointly, detecting impacts on mental health (though not on economic outcomes) only when combined.

[12] See also (Johnsen and Friborg, 2015) and (Cuijpers, Smit, et al., 2010; Cuijpers, Cristea, et al., 2016) for meta-analyses on the effectiveness of CBT in treating mental ill health.

2 The IAPT Programme

In 2008, the UK Government launched the IAPT programme to make evidence-based psychological therapies more widely available within the NHS, its universal public healthcare system, focusing on the most common mental health problems: depression and anxiety disorders.[13] At its inception, the then Secretary of State for Health and Social Care, Alan Johnson, argued: "All too often in the recent past, people experiencing anxiety and depression received relatively little help from the NHS unless their condition was particularly severe: in 2000, only 9 per cent of people [...] received psychological therapy, despite clear evidence of its effectiveness. This is something we are determined to change" (Department for Health, 2008).

What followed was an unprecedented, nationwide rollout of a mental health service, covering all 135 public health service providers (so-called *Clinical Commissioning Groups (CCGs)*, or *services* for short) in England.[14] Today, IAPT is the largest programme of its kind in the world, seen as a pioneering model for treating mental ill health at the general population level (and being replicated in other countries, e.g. Norway, Sweden, and Australia), and still growing (Clark, 2019). By now, IAPT has treated over seven million patients (more than 13% of the English population). The NHS has committed to further expand access (NHS, 2019). IAPT is a separate unit within the NHS, with its own budget.[15]

The IAPT programme provides psychological therapies recommended by the *National Institute for Health and Care Excellence (NICE)* in the UK, an independent body mandated with reviewing evidence for treatments (not limited to mental health) and issuing clinical guidelines for how effective treatments should be implemented within the NHS. For depression and anxiety disorders, NICE strongly supports psychological therapies, in particular cognitive behavioural therapy (CBT), and ad-

[13] For a detailed overview of the IAPT programme, see Clark (2018).

[14] In England, during our observation period, *Clinical Commissioning Groups (CCGs)* were independent, geographically distinct bodies accountable to the Secretary of State for Health and Social Care through NHS England, each responsible for commissioning public healthcare for, on average, about a quarter of a million of people NHS Confederation (2021). Emerging from *Primary Care Trusts (PCTs)* in 2013, CCGs were reflective of local healthcare needs. In 2022, CCGs were replaced with *Integrated Care Systems (ICS)*.

[15] There was a staggered rollout of the IAPT programme over three years: services were selected based on local demand, their capacity to supply treatments, and their willingness to opt into a mandatory IT system for collecting session-by-session patient-level outcome data. In this paper, we do not exploit this staggered rollout of the programme, which may be subject to selective buy-in as well as implementation and scaling-up effects.

vocates a stepped-care model with both low and high-intensity treatments.[16] To access the programme, patients can either be referred to by their GPs or they can refer themselves (so-called *self-referral*). The latter was a new option at the time the programme was launched that aimed to make psychological therapies more accessible amongst under-served population groups.

In their first session, patients undergo an initial assessment in which the type of problem and the severity of symptoms are determined, and in which patients and therapists jointly agree on a course of treatment. Then, patients start treatment in their second session, or, if their problem is considered more appropriate for a different service, are signposted elsewhere. In particular, those with mild to moderate symptoms start with low-intensity treatment (e.g. guided self-help, computerised CBT, or group-based physical activity programmes) and, if not responding, are upgraded to a higher intensity (e.g. usually weekly face-to-face one-to-one sessions); those with moderate to severe symptoms (as well as with special forms of anxiety disorders such as post-traumatic stress disorder) start immediately with high-intensity treatment. About 60% of patients entering the programme (over 560,000 patients per year) receive at least one clinical session. Of these, the vast majority receives treatments based on CBT, though other treatments are also available to preserve an element of choice. Overall, 30% receive low-intensity treatments based on CBT principles, 24% high-intensity CBT, 38% low-to-high-intensity stepped care (i.e. a change from low to high-intensity CBT), and 8% other forms of treatment (NHS, 2021).[17] Treatment is highly manualised. Typically, patients have one session per week, lasting around 60 minutes. There is no fixed number of sessions, though a typical course of treatment has between six and twenty sessions, with a current average of about eight sessions.

CBT refers to a wide range of psychological therapies that reduce dysfunctional emotions and behaviours by changing behaviours, appraisals of situations and thinking patterns, or both (Beck, 2020). The basic idea is that symptomatic change follows from cognitive or behavioural change, brought about by, for example, analysing maladaptive thinking patters, teaching more adaptive self-talk, or implementing more adaptive behaviours (Brewin, 1996). Take a panic attack, for

[16] See NICE Clinical Guideline 123 "Common mental health problems: identification and pathways to care" at `www.nice.org.uk/guidance/CG123`.

[17] Other forms of treatment may include, for example, interpersonal psychotherapy, couples therapy, counselling, brief psychodynamic therapy, or mindfulness-based cognitive therapy, which are recommended for depression but not for anxiety disorders.

instance: a typical CBT treatment helps patients understand what a panic attack is and how it affects them: their feelings, e.g. "I am scared"; their thinking, e.g. "I am going to pass out"; their physical symptoms, e.g. "My heart is racing and I am sweating"; and their behaviours, e.g. "I am running away from the situation". It then teaches patients to plan, implement, and, after implementation, evaluate an adaptive behavioural response, while avoiding maladaptive responses such as running away from the situation, an avoidance behaviour that eventually leads to even more panic in the future (cf. C. Williams, 2013).

Specifically for the IAPT programme, the UK Department of Health and Social Care implemented national curricula for therapists covering a wide range of evidence-based CBT treatments.[18] New therapists working in the programme are required to learn at least two treatments for depression and one for each anxiety disorder.[19] By 2019, about 10,500 new therapists were trained.

In 2018, the IAPT programme served about 17% of the community prevalence of depression and anxiety disorders, so there is more demand for psychological therapies than there is supply. This oversubscription of patients to treatments yields substantial variation in waiting times between initial assessment and start of treatment across services over time, depending on supply-side constraints (e.g. a lack of trained therapists in some services) and demand-side characteristics (e.g. clusters of mental ill health in some areas). This oversubscription and resulting exogenous variations in waiting times between sessions across services over time informs our identification strategy.

3 Data

When launched in 2008, the IAPT programme adopted an elaborate session-by-session patient-level outcome monitoring system to ensure that post-treatment outcomes are available to therapists at any point in time, even if patients finish their therapy early. This is a useful design to avoid missing endline data, which could

[18] These national curricula can be found at: https:\hee.nhs.uk. A competency framework, which specifies the clinical training and skills to deliver these treatments, can be found at: https://www.ucl.ac.uk/pals/research/cehp/research_groups/core/competence-frameworks.

[19] The training follows a joint university and on-the-job approach, whereby over a period of one year trainees attend university for several days per week (more days for trainees in high-intensity treatments, who are required to have prior experience in mental health services) and spend the rest of their time in on-the-job training.

lead to an overestimation of the effectiveness of treatment. We define a course of treatment as including the initial assessment, which is used to assess a patient's presenting problem and for determining the appropriate course of treatment, and at least two subsequent clinical sessions. As outcomes are asked *before* the start of each session, including the initial assessment, and the initial assessment has little therapeutic content, this definition allows us to track the mental health of patients from their initial assessment to at least after their first clinical session. In our sample, outcomes are available for 98% of patients who attended such a course of treatment.[20]

The protocol requires patients to complete the same clinically validated measures of depression and anxiety in each session, including the initial assessment. The procedure is that a therapist asks the patient to complete the measures in a neutral setting, on the day of the session and before the session starts, typically while patients are waiting for their appointment or earlier on the day.[21] Therapists then review these measures at the start of each session and use them for session planning. The outcome data are regularly reviewed by supervisors and service managers to ensure compliance with this protocol. While the protocol aims to avoid wasting clinical time and to reduce issues related to the self-reporting of measures (e.g. priming or demand effects), it is also a key feature of our identification strategy as it enables us to observe the evolution of mental health between initial assessment and the first clinical session without any actual treatment occurring.

Our dataset consists of the universe of patients ever treated on IAPT, entering the programme during the 2016 to 2018 period.[22] We obtain the data from NHS Digital, which include patients' session-by-session outcomes as well as rich information on their psychological-therapy and individual characteristics. We complement these patient-level data with regional data on the characteristics of services (*Clinical Commissioning Groups, CCGs*) (e.g. number of staff) from NHS Digital as well as socio-economic characteristics of local areas (e.g. local deprivation) from the Office for National Statistics (ONS) in the UK.

[20] This is in line with official statistics by NHS Digital, who report non-missing outcome data on 98.5% of patients (Digital, 2016).

[21] If treatment occurs online (e.g. via Zoom) or via phone, patients can enter their data via the internet.

[22] This covers the entire period in which the outcome monitoring system was operational, up until Covid-19.

Outcomes. Our measure for depression is the Patient Health Questionnaire 9 (PHQ-9), a routine instrument for assessing symptoms of depression amongst general and clinical populations (Kroenke, Spitzer, and J. B. W. Williams, 2001).[23] It consists of nine, four-point items that are summed up to a total, whereby scores from zero to four imply no or minimal, from five to nine mild, from ten to 14 moderate, from 15 to 19 moderately severe, and from 20 to 27 severe depressive symptoms. PHQ-9 scores equal to or greater than the clinical cut-off of ten indicate a clinical case. Our measure for anxiety is the Generalised Anxiety Disorder Questionnaire (GAD-7),[24] likewise a routine instrument for measuring anxious affect and worry (Spitzer et al., 2006). It consists of seven, four-point items that are also summed up, whereby scores from zero to four imply minimal, from five to nine mild, from ten to 14 moderate, and from 15 to 21 severe anxiety. GAD-7 scores equal to or greater than the cut-off of eight indicate a clinical case. Both measures are mandatory to collect, though therapists may also capture additional measures to assess more specific anxiety disorders.[25]

As depression and anxiety are highly co-morbid (cf. Kalin, 2020), the IAPT programme defines three main outcomes that take into account both PHQ-9 and GAD-7 scores:

1. *Reliable Improvement* is a binary indicator that is one if a patient's PHQ-9 and/or GAD-7 scores have decreased by a reliable amount and neither has shown a reliable increase.

2. *Reliable Deterioration* is, conversely, a binary indicator that is one if a patient's PHQ-9 and/or GAD-7 scores have increased by a reliable amount and neither has shown a reliable decrease.

3. *Reliable Recovery* is a binary indicator that takes on one if a patient has reliably improved *and* that patient's PHQ-9 and/or GAD-7 scores are above the

[23] The PHQ-9 asks patients about various aspects of their mood over the past two weeks and to report the frequency – ranging from "not at all" to "nearly every day" – of experiencing specific symptoms, such as how often they felt down, had little interest in doing things, felt tired, or had thoughts that they would be better off dead or of hurting themselves.

[24] The GAD-7 asks patients about their anxiety levels over the past two weeks and to report their frequency, inquiring about symptoms such as feeling nervous, not being able to stop or control worrying, worrying too much about different things, trouble relaxing, being so restless that it is hard to sit still, becoming easily annoyed or irritable, and feeling afraid, as if something awful might happen.

[25] For social anxiety disorder, for example, the Social Phobia Inventory (SPIN) (Connor et al., 2000) is collected *in addition* to both PHQ-9 and GAD-7.

clinical cut-off on either measure at the start of treatment and both are below the cut-off at the end of treatment.

IAPT uses the term *reliable* to mean a change in score that exceeds the measurement error of the scale, which for PHQ-9 is a change equal to or greater than six and for GAD-7 a change equal to or greater than four.

In defining our outcomes this way, we adopt a conservative approach that measures treatment outcomes irrespective of the specific clinical problem being treated, focusing on being free from mental ill health as the ultimate outcome of psychological therapy. As secondary outcomes on mental health, we also look at PHQ-9 and GAD-7 scores separately and at a mental health index combining both scores, standardised, as a weighted average.

We are also interested in the effect of treatment beyond measures of mental health. We look at the work and social life of patients using data from the *Work and Social Adjustment Scale* (Mundt et al., 2002), a clinically validated scale that measures patients' perceived functional impairment due to a particular health problem (here: mental ill health) overall as well as in different domains of life, including work, home management, social and private leisure, and close relationships.[26] Besides this scale, we use data on self-reported employment, in particular whether patients report to be employed as opposed to unemployed or long-term sick and whether patients report to receive statutory sick pay. As with our mental health outcomes, these are asked session-by-session. Appendix Table A.I shows summary statistics of our outcomes.

Covariates. We obtain information on patients' psychological-therapy characteristics from NHS Digital, including their referral type (e.g. whether they were referred to treatment by their GP or via self-referral), the time between referral and initial assessment in weeks, treatment mode (e.g. whether treatment was in person or online), whether they where prescribed additional medication (e.g. psychopharmacology), their initial diagnosis (e.g. whether they were initially diagnosed with depression and/or anxiety, including its type), and their treatment intensity in the

[26] The scale consists of five, eight-point items that are summed up to a total, whereby scores below ten imply no or minimal impairment, from ten to 20 significant impairment but less severe clinical symptomatology, and above 20 moderately severe or worse psychopathology. The item on work, for example, asks patients to rate: "Because of my [mental ill health], my ability to work is impaired. 0 means not at all impaired and 8 means very severely impaired to the point I can't work."

stepped-care model of the IAPT programme (e.g. whether they were in low or high-intensity treatment, or whether they changed their intensity during the course of treatment). NHS Digital also provides information on patients' individual characteristics, including their age, gender, ethnicity, religion, sexual orientation, whether they have a long-term health condition, their self-reported employment status, and whether they are a member of the armed forces. Finally, we obtain precise information on the locations and times of patients' initial assessment and all subsequent clinical sessions.

We complement these psychological-therapy and individual characteristics with regional data. In particular, to capture supply-side constraints of the IAPT programme, we obtain data on the characteristics of services (*Clinical Commissioning Groups, CCGs*), including the local number of staff, the local number of patients, and the local funding per patient, from NHS Digital. To capture demand-side characteristics, we obtain data on the socio-economic characteristics of local areas, including the local unemployment rate and median wage, likewise from NHS Digital, as well as data on local deprivation, including an index of multiple deprivation and sub-indices for deprivation in the areas of income, employment, education, health, crime, housing, and the environment, which are provided by the ONS. Appendix Table A.II shows summary statistics of our covariates.

Estimation Sample. Our sample includes all patients who started treatment between April 2016 and December 2018. We focus our analysis on this time period because certain psychological-therapy characteristics (particularly, but not limited to, the initial diagnosis) were consistently recorded only from April 2016 onwards. Moreover, according to official statistics by NHS Digital, aggregate recovery rates reached a stable level from around the same time, suggesting that the programme had moved from an initial implementation and scale-up phase to a more steady state of operation (cf. Clark, 2018), which we are primarily interested in. We remove courses of treatment that started in 2019 to keep data comparable to previous years, given the dataset discontinues at the end of 2019 and it only includes finished courses. Particularly, we do not include data on patients that started in 2019 but did not finish by the time Covid-19 pandemic disrupted data collection.

We restrict this sample to attended sessions with non-missing values for both PHQ-9 and GAD-7. We further limit ourselves to patients who are at caseness prior to treatment (i.e. who, according to their PHQ-9 or GAD-7 scores at initial assess-

ment, suffer from clinical depression or anxiety), who finished their treatment (as reported by therapists), and who completed at least three sessions (i.e. the initial assessment and at least two subsequent clinical sessions), a requirement of our research design. Our estimation sample includes 1,246,792 patients who attended, on average, 7.7 sessions (standard deviation of 4.1).[27]

4 Empirical Strategy

4.1 Identification

Our aim is to estimate the causal effect of being treated within the IAPT programme. In the potential outcomes framework by Rubin (1974), the average treatment effect on the treated (ATT) can be written as the average difference in the outcomes between patients who receive treatment and those who do not.

Suppose patient i's initial assessment was at time t_i and the duration of the (potential) treatment is w. For the moment, take w as fixed and suppose we only consider a subset of the data for these patients. Let t_{i1} and t_{i2} respectively denote t_i and $t_i + w$. We introduce the following variables: D_{it_i} is the treatment dummy that takes value 1 for the treated; $Y_{it_{ij}}(0)$ is the outcome for patient i at time t_{ij} if they were to not receive treatment; and $Y_{it_{ij}}(1)$ is the outcome for patient i at time t_{ij} if they were to receive treatment.

Our parameters of interest are ATT and CATT (conditional ATT) that we denote respectively by θ and $\theta(X_{it_i})$. They are formally defined as follows,

$$\begin{aligned} \theta &:= E\left[Y_{it_{i2}}(1) - Y_{it_{i2}}(0) \,|D_{it_i} = 1\right], \\ \theta(X_{it_i}) &:= E\left[Y_{it_{i2}}(1) - Y_{it_{i2}}(0) \,|D_{it_i} = 1, X_{it_i}\right]. \end{aligned}$$

ATT and CATT are not identified without further assumptions since we only observe $Y_{it_{ij}} := D_{it_i} Y_{it_{ij}}(1) + (1 - D_{it_i}) Y_{it_{ij}}(0)$, but never both $Y_{it_{ij}}(1)$ and $Y_{it_{ij}}(0)$. The identifying assumptions we make are standard in the econometrics literature on difference-in-differences models when two time periods are available, as recently surveyed by J. Roth et al. (2023). To increase the credibility of the assump-

[27] When cross-validating the properties of our estimation sample with official statistics by NHS Digital, we find a very similar recovery rate: 55.5% in our sample vs. 49.3% (NHS, 2017). Recall that, given our research design, we calculate recovery rates from a course of treatment that includes at least three sessions. The NHS defines a course of treatment as including at least two sessions.

tions, these are made conditional on X_{it_i} that represents a vector of observed characteristics associated with patient i. It is convenient to define $\Delta Y_{it_i} := Y_{it_{i2}} - Y_{it_{i1}}$ and $\Delta Y_{it_i}(d) := Y_{it_{i2}}(d) - Y_{it_{i1}}(d)$ for $d = 0, 1$. We assume the following assumptions hold throughout.

Assumption 1: Parallel trends. For all i,

$$E\left[\Delta Y_{it_i}(0) | D_{it_i} = 1, X_{it_i}\right] = E\left[\Delta Y_{it_i}(0) | D_{it_i} = 0, X_{it_i}\right] \text{ almost surely.}$$

Assumption 2: No anticipatory effects. For all i,

$$E\left[Y_{it_{i1}}(0) | D_{it_i} = 1, X_{it_i}\right] = E\left[Y_{it_{i1}}(1) | D_{it_i} = 1, X_{it_i}\right] \text{ almost surely.}$$

In our context, Assumption 1 states that the expected natural recovery for patients in the treatment and control group are the same if there were no IAPT programme. Assumption 2 states that the expected initial outcome, before any treatment, for patients in the treatment group is not affected by them being in the treatment group.

Under Assumptions 1 and 2, the observed change in expected outcomes for the treatment group can be decomposed into the treatment effect and the observed change in expected outcomes for the control group. That is, we can write ATT and CATT in the difference-in-differences form for observables, namely:

$$\theta = E\left[\Delta Y_{it_i} | D_{it_i} = 1\right] - E\left[\Delta Y_{it_i} | D_{it_i} = 0\right], \tag{1}$$

$$\theta(X_{it_i}) = E\left[\Delta Y_{it_i} | D_{it_i} = 1, X_{it_i}\right] - E\left[\Delta Y_{it_i} | D_{it_i} = 0, X_{it_i}\right]. \tag{2}$$

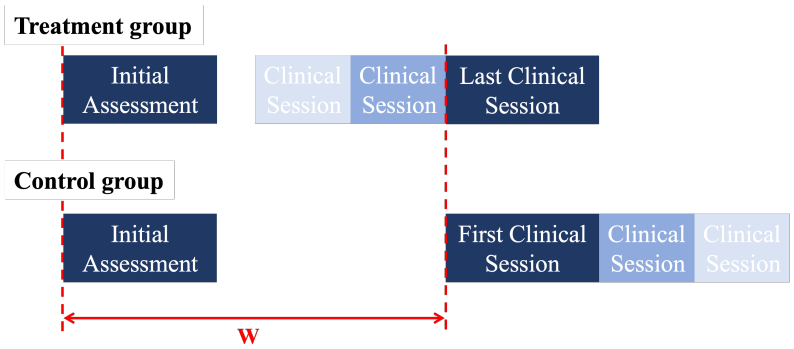
In Appendix B, we provide a proof that ATT and CATT can be written in terms of the distribution of observables, along with related discussions.

We analyse our data through the lens of a two-period model, which is justified under the assumption that $\{(\Delta Y_{it_i}, D_{it_i}, X_{it_i})\}_{i=1}^{n}$ is a random sample that in turn imposes the stable unit treatment value assumption and stationarity of the data generating process. It is worth emphasising that our patients enter the programme at different times (hence the t_i subscript), so that our data are not suitable to be studied under a multi-period, cohort-wide adoption of a staggered-treatments framework, which is the main focus in the survey by J. Roth et al. (2023).

A well designed and carefully executed RCT can ensure that Assumptions 1 and 2 hold. However, the IAPT programme has not been implemented as an RCT.

We thus take a quasi-experimental approach and argue that Assumptions 1 and 2 reasonably hold. We do so by exploiting the oversubscription of patients to the programme and resulting exogenous variations in waiting times between sessions across services over time for identification. In particular, we create a quasi-experimental control group using patients who, after their initial assessment, are waiting for their first clinical session. We then compare the change in mental health outcomes for patients between their initial assessment and their last clinical session (our treatment group) with the change in mental health outcomes for patients between their initial assessment and their first clinical session (our control group). In doing so, we are comparing patients who reach respective sessions (the last clinical session for our treatment group, the first for our control group) around the same time after initial assessment. Figure 1 illustrates our research design.

Figure 1: Research Design – Waitlist-Based Quasi-Randomisation



Note: Own illustration.

Given that X_{it_i} includes psychological-therapy, individual, and local-area characteristics, as well as fixed differences between services and time fixed effects, we believe that Assumptions 1 and 2 reasonably hold. Note that Assumption 1 is weaker than assuming that treatment assignment in our quasi-experiment is random conditional on X_{it_i}.

Before moving on to estimation, we provide further discussions to support the validity of Assumption 1 that may even extend to the stronger restriction that the treatment assignment in our quasi-experiment can be treated as random conditional on X_{it_i}.

Selection. While controlling for psychological-therapy, individual, and local-area characteristics, as well as fixed differences between services and time fixed effects

ensures that waiting times are conditionally independent from outcomes, there may be concern about residual selection.

When it comes to *within-sample selection*, there may be a concern that therapists could prioritise patients with worse mental health, or certain demographics. This is avoided due to the stepped-care protocol of the IAPT programme: after referral and initial assessment, therapists allocate patients to either low or high-intensity treatment, in each of which they are processed. This allocation is done on a first-come first-serve basis, based on fairness principles, and is rigorously followed through by therapists.[28] In line with this, we do not observe a significant correlation between waiting time and either PHQ-9 or GAD-7 scores, which would be indicative of prioritisation of patients.[29]

Appendix Table A.III shows balancing properties of covariates between our treatment and our quasi-experimental control group, which uses the 50th percentile of waiting time (between 21 and 37 days, depending on the intensity of treatment) as a default threshold. Following Imbens and Rubin (2015), we calculate four scale-free overlap measures: normalised differences (which, unlike simple differences in means and associated t-tests, are insensitive to the number of observations) and, to measure dispersion of covariates between groups, the logs of the ratios of standard deviations and the shares of the control (treated) units outside the 0.025 and 0.975 quantiles of the covariate distribution of the treated (control) units. As seen, almost none of the normalised differences exceeds 0.25, which Imbens and Wooldridge (2009) suggest as a threshold above which covariates can be considered unbalanced. The only noticeable imbalance is that a larger share of the treated are treated via phone (and, in turn, a smaller share face-to-face). Moreover, there are almost no noticeable differences in dispersion of covariates between groups, as indicated by logs of the ratios of standard deviations that are below one and shares of the units outside the 0.025 and 0.975 quantiles of the counterpart covariate distribution that are close to zero. Covariates are, therefore, well balanced between groups and our treatment and quasi-experimental control group well comparable.

Appendix Table A.IV replicates Table A.III, by showing balancing properties of outcomes between our treatment and our quasi-experimental control at the start of different sessions. As seen, neither at initial assessment nor at the start of the first

[28] If present, prioritisation would lead to a lower-bound estimate. To the extent that the initial assessment itself has a therapeutic value, this does not bias our results as it is balanced between groups.

[29] That is, $r = 0.017$ for PHQ-9 and $r = 0.016$ for GAD-7.

or last clinical sessions does any of the normalised differences exceed the recommended threshold of 0.25 (Imbens and Wooldridge, 2009). There is little evidence for an unusual dispersion of outcomes between groups at any point in time either. Patients in our treatment and our quasi-experimental control group are, therefore, well comparable in terms of outcomes at the start of therapy and after therapy has ended, as well as when attending their first clinical session.

When it comes to *out-of-sample selection*, a potential issue may arise with patients discontinuing treatment. If attrition is selective – meaning that the probability of dropping out is correlated with the likelihood of recovery – it may introduce bias into our treatment effect estimates. For example, patients in our quasi-experimental control group may naturally recover during the wait between initial assessment and their first clinical session and, therefore, drop out of the programme. To reduce this concern, in Appendix Section D, we establish bounds on our treatment effect estimates by imputing outcomes under various scenarios. Even under the most extreme assumptions, such as that all dropped-out respondents who would have been assigned to the treatment group and experience deterioration and all those dropping-out of the control group experience recovery, our estimated treatment effects for both reliable recovery and reliable improvement remain significant and positive.

Similarly, individuals in our quasi-experimental control group may, during the wait between initial assessment and their first clinical session, opt for an alternative treatment outside of the IAPT programme while still being part of the programme. This would introduce upward bias in natural recovery, suggesting that our estimated treatment effects can be interpreted as a lower bound. Note that the IAPT programme is run by the NHS, which is the monopolist provider of state-funded healthcare in England.

Waitlist Effects. Finally, there may be concern that waiting itself could have a negative impact and, thereby, introduce downward bias in natural recovery. We argue that this is unlikely to be strong enough to have a direct effect on mental health. The reason is that waiting is to be expected by all patients. Criticisms on waiting times in the NHS have long been well-publicised, so that having to wait is common knowledge. Moreover, the first-come first-serve principle and associated waiting times are announced at initial assessment. Empirically, Appendix Figure A.I plots our main outcomes – reliable recovery, improvement, and deterioration – as raw

data for different waiting times. As seen, there is evidence for a slight natural recovery, which is, however, quantitatively minor. We see that the states of waitlisted patients are more likely to improve than to deteriorate. Hence, our estimated treatment effects come from the intervention being beneficial, rather than from the wait being detrimental.

4.2 Estimation

4.2.1 Average Treatment Effects

In Section 4.1, we only consider patients that have w weeks as the duration of or waiting time for treatment. We now combine observations for many w's and update our notation by letting $\Delta Y_{it_i} := D_{it_i}\Delta Y^{tr}_{it_i} + (1 - D_{it_i})\,\Delta Y^{c}_{it_i}$, with $\Delta Y^{tr}_{it_i} := Y_{it_i+W_{it_i}}(1) - Y_{it_{i1}}(1)$ and $\Delta Y^{c}_{it_i} := Y_{it_i+W_{it_i}}(0) - Y_{it_{i1}}(0)$. That is, ΔY_{it_i} is the change in the outcome of individual i, which is the change between initial assessment and the last clinical session if i belongs to our treatment group, $\Delta Y^{tr}_{it_i}$; and the change between initial assessment and the first clinical session if i belongs to our control group, $\Delta Y^{c}_{it_i}$, cf. Figure 1. W_{it_i} denotes the duration of or waiting time for treatment respectively for a patient in the treatment or control group. D_{it_i} is the treatment dummy, which is one if i's first clinical session falls below a pre-defined threshold of waiting time. Our default threshold is the 50th percentile, which is between 21 and 37 days, depending on the intensity of treatment.[30]

In addition to Assumptions 1 and 2, let us suppose that the following holds:

$$E\left[\Delta Y_{it_i} | D_{it_i}, W_{it_i}, X_{it_i}\right] = \beta_0 + \beta_1 D_{it_i} + \beta_2 W_{it_i} + \beta_3^{\top}\tilde{X}_{it_i} + \mu_{ir} + \nu_{it_i}. \quad (3)$$

Then, β_1 represents the ATT. Here, X_{it_i} are psychological-therapy and individual characteristics, measured prior to treatment; $\tilde{X}_{it_i}$ are service and local-area characteristics; and μ_{ir} and ν_{it_i} are service (i.e. 135 CCGs) and time fixed-effects (i.e. day-of-week, month, and year), respectively. We also routinely control for waiting time (on average and specific to each intensity of treatment) and time lapsed between referral and initial assessment in weeks as well as for pre-treatment mental health (in form of our standardised mental health index) throughout our regressions.

[30] The threshold is 24 days for low and 21 days for high-intensity treatment, 32 days for stepped-up courses, 37 days for stepped-down courses, and 28 days if the treatment intensity is undefined (due to multiple changes).

We estimate the following linear model:

$$\Delta Y_{it_i} = \beta_0 + \beta_1 D_{it_i} + \beta_2 W_{it_i} + \beta_3^\top \tilde{X}_{it_i} + \mu_{ir} + \nu_{it_i} + u_{it_i}. \tag{4}$$

Note that the time-varying covariates net systematic differences between our treatment and control group at the psychological-therapy and individual level as well as at the service and local-area level (e.g. differences in local deprivation over time that may be directly related to our outcome and, indirectly via waiting time, to our treatment dummy), whereas the service and time fixed effects net out any remaining unobserved heterogeneity between services over time. We estimate treatment effects in Equation 4 using OLS with robust standard errors clustered at the service level.[31]

4.2.2 Heterogeneous Treatment Effects

Under Equation 3, the treatment effect is expected to be the same for all types of patients. We now estimate how the effectiveness of the IAPT programme varies across patients, services, and areas with different characteristics. We take two approaches. First, we construct matching estimators using a pre-selected set of previously observed sources of heterogeneity, as found in earlier literature based on reduced-form analysis of treatment outcomes. Second, we use a state-of-the-art machine learning (ML) technique and let the data tell us the most relevant sources of heterogeneity for the treatment effect. Specifically, for the latter, we use the *generalised random forest*, a data-driven way to identify the sources of heterogeneity amongst all available covariates. The validity of our estimators in terms of identifying the treatment effect follows under the same assumptions as outlined in Section 4.1.

ATT with pre-selected sources of heterogeneity. We are interested in whether the treatment effect differs for different patients, services, and areas, and if so, what characteristics are associated with better or worse outcomes. Using a similar notation as before, let our data be $\{(\Delta Y_{it_i}, D_{it_i}, W_{it_i}, Q_{it_i})\}_{i=1}^n$. To facilitate matching, we dichotomise the covariates that have been shown in earlier literature to be related to heterogeneity in treatment outcomes and enumerate each combination as a patient type. We use Q_{it_i} to represent the type indicator that each patient belongs

[31] Given that ΔY_{it_i} is discrete for out main outcomes, in Section 5.2 we provide the results of logit model as a robustness check.

to. Our CATT is then indexed by (w, q), which corresponds to a particular treatment/waiting time duration and patient type. In this case, under Assumptions 1 and 2, our CATT can be written for each (w, q) as (cf. Equation 2),

$$\theta(w, q) := E[\Delta Y^{tr}_{it_i} | W_{it_i} = w, Q_{it_i} = q] - E[\Delta Y^{c}_{it_i} | W_{it_i} = w, Q_{it_i} = q]. \quad (5)$$

Since (W_{it_i}, Q_{it_i}) are discrete, there are finite combinations of (w, q). We can estimate $\theta(w, q)$ nonparametrically by calculating the difference between the average outcomes of the treated and the control group patients whose $W_{it_i} = w$ and $Q_{it_i} = q$. We only include sub-populations that have a sufficient number of observations for both treatment and control group.[32] Sub-populations that have too few observations and those that do not have a treatment or control group counterpart are excluded from the analysis. This implies that we only use the treated patients that have a close control-group counterpart, and *vice versa*.

Stacking the nonparametric estimators for $\theta(w, q)$ over (w, q) gives us a vector of CATTs that has an asymptotically normal distribution following from a standard central limit theorem. Furthermore, the asymptotic distribution of the vector of CATTs can be consistently bootstrapped using the standard resampling method with replacement since the empirical measure can be bootstrapped in this way (Gine and Zinn, 1990). Conveniently, however, the nonparametric estimator just described is numerically equivalent to the OLS estimator of $\{\theta(w, q)\}$ from this saturated model:

$$\begin{aligned} \Delta Y_{it_i} \quad = \quad & \sum_{w,q} \beta(w, q) \times \mathbf{1}\{Q_{it_i} = q, W_{it_i} = w\} \\ & + \sum_{w,q} \theta(w, q) \times \mathbf{1}\{Q_{it_i} = q, W_{it_i} = w\} \times D_{it_i} + u_{it_i}. \end{aligned} \quad (6)$$

where $\mathbf{1}\{Q_{it_i} = q, W_{it_i} = w\}$ is a dummy which is one if the patient was either treated in or waited for w weeks and belongs to type q. We provide a proof of this equivalence in Appendix B. Thus, in practice, we use the above linear equation to estimate the CATTs by OLS, which provides a simple framework for inference on $\{\theta(w, q)\}$. That is, one can readily test the homogeneity hypothesis on the CATTs, where the null hypothesis states that all CATTs are equal, using a Wald test.

[32] The results are reported for a minimum of 100 observations per treatment and control group.

ATT with data-driven sources of heterogeneity. To further explore heterogeneities in the treatment effect without constraining the analysis to a set of preselected sources, we use the *generalised random forest* (Athey, Tibshirani, and Wager, 2019).

The basic idea behind this algorithm is to split the sample into bins that share the same realisations of covariates. Then, the observations in the bin are used to estimate a bin-specific treatment effect. The partition into bins is performed to maximise the heterogeneity in within-bin treatment effect estimates across bins. Hence, by design, the algorithm searches for variables that are related to differences in treatment effects. The procedure is then repeated across many subsamples and the treatment effect estimates are averaged to reduce variance.[33]

To take it to a more familiar context, a forest can be thought of as a nearest-neighbor method, in that it performs the estimation using a weighted average of observations in the "neighborhood". However, in contrast with classical methods, the neighborhood is defined in a data-driven way. The advantage of this approach is that it defines a neighborhood in a flexible way depending on the data at hand. By treating the forest as an adaptive nearest-neighbor estimator, Athey, Tibshirani, and Wager (2019) show that the estimates of the generalized random forest are consistent and asymptotically normal.

We find that being treated within the IAPT programme significantly improves patients' mental health outcomes. In particular, it increases the likelihood to reliably recover by about 43 and to reliably improve by about 38 percentage points, on average, while reducing the likelihood to deteriorate by about 8 percentage points. The latter suggests that the programme has, on average, no adverse effects, which is a contribution in its own right addressing recent concerns that well-intended psychological interventions can have unintended consequences (cf. Harvey et al., 2023). Point estimates and associated standard errors are remarkably similar in size regardless of whether we include covariates or not.

[33] In practice, the algorithm uses different subsamples for binning and treatment effect estimation. This is known as the *honest* approach that serves to avoid overfitting and biasing estimates. As a technical note, we assume that potential outcomes are independent of treatment assignment, conditional on the set of covariates. Our algorithm incorporates this conditioning by orthogonalising the treatment indicator and the outcomes and calculating the within-bin treatment effect estimate from regression residualised outcomes on residualised propensity scores. This technique is sometimes known as *double machine learning*, which is particularly important for our application given that we use observational rather than experimental data. For further details on double machine learning, see Chernozhukov et al. (2018).

We refer the reader to Athey, Tibshirani, and Wager (2019) and Wager and Athey (2018) for the detailed account of the algorithm and its corresponding asymptotic theory.

5 Results

5.1 Average Treatment Effects

Table 1 shows the average treatment effects on our main outcomes – reliable recovery, improvement, and deterioration – using our default control group (50th percentile of waiting time). Columns 1, 3, and 5 show models without controls, Columns 2, 4, and 6 models that control for psychological-therapy, individual, service, and local-area characteristics as well as service (i.e. 135 CCGs) and time fixed-effects (i.e. day-of-week, month, and year), which are our preferred models. Recall that our dependent variables are binary, and that we are estimating linear probability models.

Table 1: Average Treatment Effects on Mental Health

	Reliable Recovery (0-1)		Reliable Improvement (0-1)		Reliable Deterioration (0-1)	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.443***	0.431***	0.388***	0.377***	-0.085***	-0.084***
	(0.004)	(0.004)	(0.004)	(0.003)	(0.002)	(0.001)
Therapy Controls	No	Yes	No	Yes	No	Yes
Individual Controls	No	Yes	No	Yes	No	Yes
Service Controls	No	Yes	No	Yes	No	Yes
Local-Area Controls	No	Yes	No	Yes	No	Yes
Service Fixed Effects	No	Yes	No	Yes	No	Yes
Time Fixed Effects	No	Yes	No	Yes	No	Yes
Number of Individuals	1,246,792	1,246,792	1,246,792	1,246,792	1,246,792	1,246,792
Treatment Group	618,574	618,574	618,574	618,574	618,574	618,574
Control Group	628,218	628,218	628,218	628,218	628,218	628,218
R Squared	0.228	0.289	0.152	0.187	0.022	0.064

Note: Linear probability models. Binary dependent variables. Robust standard errors clustered at service level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Next, Table 2 presents the results of the main streams of the IAPT programme's stepped-care model, by splitting Table 1 into its different treatment intensities. Panel A shows the average treatment effects for patients in the low-intensity treatment, Panel B for patients in the high-intensity treatment, and Panel C for patients who are stepped up from initially low to then high-intensity. The full results, which include smaller streams (e.g. patients who are stepped down from initially high to then low-intensity treatment or patients for whom the intensity was not recorded), are presented in Appendix Table C.I.

Table 2: Average Treatment Effects on Mental Health by Treatment Intensity

	Reliable Recovery (0-1)		Reliable Improvement (0-1)		Reliable Deterioration (0-1)	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Low Intensity						
Treatment	0.440***	0.430***	0.368***	0.360***	-0.078***	-0.078***
	(0.005)	(0.005)	(0.004)	(0.004)	(0.002)	(0.002)
Number of Individuals	491,942	491,942	491,942	491,942	491,942	491,942
Treatment Group	245,433	245,433	245,433	245,433	245,433	245,433
Control Group	246,509	246,509	246,509	246,509	246,509	246,509
R Squared	0.216	0.284	0.138	0.179	0.020	0.053
Panel B: High Intensity						
Treatment	0.439***	0.429***	0.404***	0.393***	-0.084***	-0.084***
	(0.008)	(0.008)	(0.007)	(0.006)	(0.003)	(0.002)
Number of Individuals	275,990	275,990	275,990	275,990	275990	275990
Treatment Group	136,379	136,379	136,379	136,379	136379	136379
Control Group	139,611	139,611	139,611	139,611	139611	139611
R Squared	0.234	0.298	0.164	0.198	0.021	0.069
Panel C: Step Up (Low to High Intensity)						
Treatment	0.449***	0.435***	0.404***	0.385***	-0.095***	-0.090***
	(0.004)	(0.005)	(0.004)	(0.004)	(0.002)	(0.002)
Number of Individuals	388,136	388,136	388,136	388,136	388136	388136
Treatment Group	191,868	191,868	191,868	191,868	191868	191868
Control Group	196,268	196,268	196,268	196,268	196268	196268
R Squared	0.244	0.296	0.164	0.200	0.024	0.078
Therapy Controls	No	Yes	No	Yes	No	Yes
Individual Controls	No	Yes	No	Yes	No	Yes

Service Controls	No	Yes	No	Yes	No	Yes
Local-Area Controls	No	Yes	No	Yes	No	Yes
Service Fixed Effects	No	Yes	No	Yes	No	Yes
Time Fixed Effects	No	Yes	No	Yes	No	Yes

Note: Linear probability models. Binary dependent variables. Robust standard errors clustered at service level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

In line with our previous results, we find that treatment significantly increases the likelihood to reliably recover and improve while decreasing the likelihood to deteriorate in each treatment intensity, by about the same size. Similar impacts across treatment intensities suggest that the allocation of patients by therapists to different treatment intensities results in an appropriate patient-therapy fit.

One might be tempted to think that different treatment intensities are redundant if these lead to similar treatment effects. Note, however, that patients in different treatment intensities are different and have different therapeutic needs. Appendix Table C replicates Table 2 by replacing our main outcomes – reliable recovery, improvement, and deterioration – with changes in underlying PHQ-9 and GAD-7 scores as well as changes in our mental health index. Recall that reductions in these outcomes imply improvements in mental health. As seen, patients in the high-intensity treatment show much stronger reductions in their PHQ-9 and GAD-7 scores as well as in our mental health index, and so do patients for whom treatment intensity is changed during their course of treatment. This suggests that different treatment intensities do indeed cater to different needs, which is also reflected in differences in underlying therapies and mechanisms, as outlined in Section 2.

Next, we exploit our session-by-session outcome data to look at the value added of different clinical sessions. Appendix Figure A.II shows reliable recovery for different bins of sessions, separately for patients who have a total of three, seven, nine, and 13 sessions, equivalent to the 25th, 50th, 75th, and 90th percentile in the overall session distribution. For example, *Sessions 5* for patients who have a total of nine sessions is the value added, in terms of reliable recovery, of having attended five out of the nine sessions, while *Sessions 9* is the value added of having attended all sessions. To reduce heterogeneity, in each case, the control group is restricted to patients who have the same number of total sessions. We make two observations: first, the relative session value added is lower for patients who have a higher total number of sessions (and who are, presumably, sicker). For example, the value added

of having attended five sessions is only nine percentage points for patients who have a total of 13 sessions, yet 14 percentage points for those who have a total of nine and even 22 percentage points for those who have a total of seven sessions. That is, the rate of improvement from mental ill health is lower the higher the number of total sessions. Second, most of the session value added, in terms of reliable recovery, is generated during the last two sessions, regardless of the total number of sessions. For these last two sessions, causality may also go the other way around: therapists may discard patients after they have reached a particular threshold of recovery, and patients leave the programme. Appendix Figures A.III and A.IV replicate Figure **??** for reliable improvement and deterioration, showing a similar pattern for the relative session value added (though with a more equal value added of different bins). Finally, we look at ripple effects of improved mental health on patients' work and social life. We do so in two ways: first, we look at changes in the *Work and Social Adjustment Scale* (Mundt et al., 2002), a clinically validated scale that measures patients' perceived functional impairment due to a particular health problem (here: mental ill health) overall as well as in different domains of life. Second, we look at changes in employment as a result of treatment. We are particularly interested in patients who report being unemployed, being long-term sick, or receiving statutory sick pay at the start of treatment, and hence look at the change from being unemployed to being employed, from being long-term sick to being employed, and from receiving statutory sick pay to not. As with mental health, these outcomes have been recorded on a session-by-session basis.

Appendix Table C.V shows our average treatment effects on the *Work and Social Adjustment Scale*. As seen, being treated within the IAPT programme significantly and strongly reduces patients' perceived functional impairment due to mental ill health, decreasing overall impairment by 5.7 points on a 0-to-40 scale (65% SD of the pre-treatment score in the treatment group), driven in almost equal parts by reductions in each domain of life (each between one and 1.4 points on a 0-to-8 scale), including work (-1.1 points, 42% SD of the pre-treatment score). That is, patients who undergo psychological therapy report to function better in all domains of life afterwards.

Appendix Table C.VI shows our average treatment effects on employment as a result of treatment.[34] As seen, being treated within the IAPT programme has, over-

[34]Different from our previous analysis, we estimate treatment effects by regressing post-treatment employment on pre-treatment employment and our treatment dummy, all other things

all, no or only negligible effects on employment. However, when restricting our sample to patients who were unemployed or long-term sick at the start of treatment, we find that being treated significantly increases their likelihood to be employed by three and two percentage points, respectively, while decreasing their likelihood to receive statutory sick pay by three percentage points. Although these effects are small, they are very short-term, as employment is last measured at the beginning of the last clinical session, and the typical course of treatment lasts between six to twenty weeks. That is, there is some evidence for small, positive short-term impacts on employment of patients who undergo psychological therapy.

5.2 Robustness Checks

We conduct a range of robustness checks for our average treatment effects obtained from estimating Equation 4.

Our results are robust to different definitions of treatment and control group when varying treatment and corresponding waiting time durations. Appendix Table C.III uses, instead of the 50th percentile of waiting time, the 25th, 75th, and 90th percentile, respectively, to allocate patients into treatment and control group. As seen, our estimates are qualitatively the same as before.

Our results are also robust to different models, samples, and outcomes. Appendix Table C.IV Column 1 estimates a logit instead of a linear probability model. Columns 2 and 3 selectively exclude certain mental health problems: Column 2 excludes patients who have substance abuse disorders as these exhibit different behaviours when on a waitlist than others (J. Williams and Bretterville-Jensen, 2022), whereas Column 3 focuses only on patients who have depression and anxiety disorders, the main target population of the IAPT programme. Finally, Columns 4 to 6 replace our main outcomes – reliable recovery, improvement, and deterioration – with changes in PHQ-9 and GAD-7 scores as well as changes in our mental health index. As seen, in all cases, our results remain robust.

We address potential concerns about attrition in Appendix Section D, where we show that the programme remains effective even under extreme assumptions

being the same. This is because patients can be either employed or not, respectively, at the start and at the end of treatment, which may, when switching from employed to not employed, result in a difference in our employment outcome of minus one, which cannot be estimated using a linear probability model. We circumvent this issue using an alternative value-added model. Note that all of our previous results continue to hold when using this alternative model.

on the outcomes of patients who discontinued treatment.

5.3 Heterogeneous Treatment Effects

We now focus on the CATT estimates of our main outcomes: reliable recovery, reliable improvement, and reliable deterioration. The CATT estimates presented here are based on our default control group, which uses the 50th percentile of waiting time.

Results with pre-selected sources of heterogeneity. Figure 2 presents the histograms of our heterogeneous treatment effect estimates produced by the matching approach described in Section 4.2.2. The vertical dashed line represents the estimated average treatment effect.[35] We selected potential sources of heterogeneity based on earlier findings on characteristics correlated with treatment outcomes and include treatment intensity, severity of the symptoms at the initial assessment, ethnicity, religion, presence of a long-term health condition, service size, and funding as well as area deprivation.[36] [37]

We find statistically significant heterogeneity in the treatment effect across sub-populations. By studying the sub-populations with the lowest and the highest treatment effects, we show that, although the programme increases the probability of recovery and improvement for all sub-populations of patients considered, there are some for whom the programme does *not* decrease the probability of deterioration.

[35] The estimators described in Section 4.2.2 can also be used to estimate the ATT by aggregating CATTs. These average effects, both from using pre-selected or data-driven observed heterogeneities, are in line with the results of the ATT estimates presented in Section 5.1. The nonparametric matching approach estimates the ATT of the programme to be 0.434 (0.001) for reliable recovery, 0.379 (0.001) for reliable improvement, and -0.086 (0.001) for reliable deterioration. In the machine-learning approach, the ATT is estimated to be 0.436 (0.001) for reliable recovery, 0.383 (0.001) for reliable improvement, and -0.089 (0.001) for reliable deterioration.

[36] The covariates are selected based on the following earlier studies. Gyani et al. (2013): course intensity, a binary indicator for severity of symptoms above the median at initial assessment, and severity as a z-score constructed from PHQ-9 and GAD-7 scores at initial assessment; Moller et al. (2019): ethnicity, religion, and presence of a long-term health condition; Clark (2018) and Gyani et al. (2013): binary indicators for service size by number of staff and service funding per patient above the median; Jaime Delgadillo et al. (2016): a binary indicator for area deprivation above the median.

[37] After eliminating observations that do not have a match, we are left with 76% of the original sample or 947,457 observations spread over 1,171 matched sub-populations. The summary statistics of the outcomes and covariates in the original and the final sample are presented in Appendix Table E.I. The sub-populations are well-balanced in terms of the number of treated and control observations. The share of treated observations varies from 22% to 82% with an average of 49%.

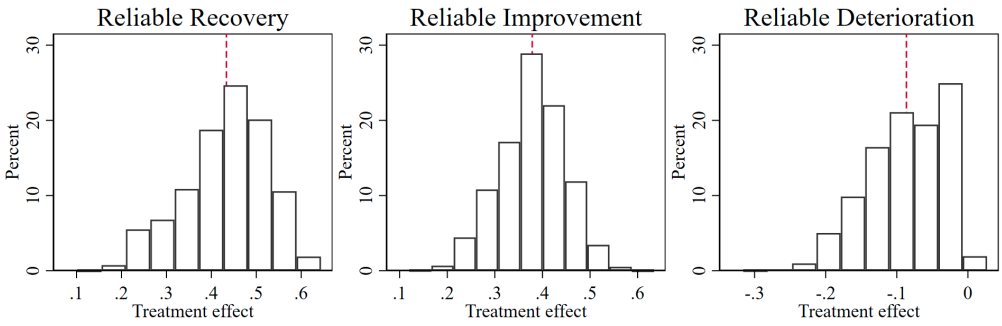


Figure 2: Conditional Average Treatment Effects – Matching Approach

Note: The histograms plot the distributions of conditional average treatment effects, which are estimated as a difference in average outcomes between treatment and control group observations in sub-populations formed by combinations of psychological-therapy, individual, service, and local area characteristics. The estimates are weighted by the number of treatment-group observations in each sub-population.

To understand in more detail which specific characteristics are systematically associated with better or worse treatment effects, we estimate the following model:

$$\begin{aligned}\Delta Y_{it_i} = \beta_0 + \beta_1 D_{it_i} + \sum_q \beta_q Q_{it_i} + \sum_w \beta_w W_{it_i} \\ + \sum_q \gamma_q Q_{it_i} D_{it_i} + \sum_w \gamma_w W_{it_i} D_{it_i} + u_{it_i},\end{aligned} \tag{7}$$

where, to assess how the effect of the treatment differs for different sub-populations, the treatment dummy, D_{it_i}, is interacted with the psychological-therapy and individual as well as service and local-area characteristics, Q_{it_i}. γ_q in Equation 7 is informative on how treatment effects vary for different patients. Table 5.3 presents the estimates of the coefficients on the interaction between these characteristics and the treatment dummy. The full results are presented in Appendix Table E.II.

We find moderate heterogeneity in treatment effects across different intensities of treatment, with patients in high-intensity treatments being more likely to reliably improve and less likely to reliably deteriorate. Importantly, patients were four percentage points less likely to reliably recover and seven percentage points less likely to reliably improve if they were provided with a treatment that was *not* recognised high or low intensity, e.g. treatments that were not on the list of NICE-recommended treatments for a given diagnosis. This result replicates the finding from Gyani et al. (2013)) which is based on the comparison of pre- and post-

treatment scores, and highlights that compliance with NICE recommendations is strictly associated with better patient outcomes.

Table 3: Heterogeneous Treatment Effects on Mental Health

	Reliable recovery	Reliable improvement	Reliable deterioration
Course intensity: Low intensity # Treated	0	0	0
	(.)	(.)	(.)
High intensity # Treated	0.002	0.039***	-0.016***
	(0.002)	(0.003)	(0.002)
Step down # Treated	0.003	0.017	0.001
	(0.010)	(0.012)	(0.007)
Step up # Treated	-0.018***	0.021***	-0.019***
	(0.002)	(0.003)	(0.002)
Undefined # Treated	-0.036***	-0.066***	-0.011
	(0.012)	(0.013)	(0.008)
Severity above median # Treated	-0.088***	-0.071***	0.096***
	(0.002)	(0.002)	(0.001)
Deprivation above median # Treated	-0.027***	0.004**	-0.014***
	(0.002)	(0.002)	(0.001)
Long-term health condition # Treated	-0.026***	0.003	-0.008***
	(0.003)	(0.003)	(0.002)
Service size above median (number of staff) # Treated	-0.004**	-0.006***	0.003**
	(0.002)	(0.002)	(0.001)
Service funding per patient above median # Treated	0.021***	0.026***	-0.010***
	(0.002)	(0.002)	(0.001)
Religion: Christian # Treated	0	0	0
	(.)	(.)	(.)
Not religious # Treated	-0.025***	-0.013***	0.007***
	(0.003)	(0.003)	(0.002)
Other religion and missing # Treated	-0.030***	-0.021***	0.006***
	(0.003)	(0.004)	(0.002)
Ethnicity: White British # Treated	0	0	0
	(.)	(.)	(.)
Other # Treated	-0.018**	0.000	-0.016***
	(0.007)	(0.008)	(0.005)
Missing # Treated	-0.055***	-0.030***	0.002
	(0.003)	(0.003)	(0.002)
R2	0.26	0.16	0.05
Observations	947,547	947,547	947,547

Note: Linear probability models. Binary dependent variables. Robust standard errors clustered at service level in parentheses. The full results are presented in Appendix Table E.II.
*** p<0.01, ** p<0.05, * p<0.1.

We also find that patients with higher severity at the beginning of treatment are less likely to reliably recover. This is perhaps not surprising, given that patients with more severe symptoms need to show considerably more improvement to be classified as reliably recovered. We see that patients with higher severity are also less likely to reliably improve and more likely to deteriorate.

In terms of heterogeneity by patient characteristics, patients with long-term health conditions are around three percentage points less likely to reliably recover. The direction of the gap confirms findings by Moller et al. (2019) for the difference in treatment outcomes. However, the difference in outcomes found by Moller et al. (2019) is significantly higher in magnitude, at 14 percentage points. This likely indicates that a large part of the difference estimated by Moller et al. (2019) is due to the difference in natural recovery rates. We also find that non-White-British patients, or those whose ethnicity is not recorded, perhaps reflecting the data collection quality, are less likely to reliably recover. Non-religious patients are less likely to reliably recover or improve and more likely to deteriorate compared to patients who identify as Christian.

For area characteristics, patients in more deprived areas are less likely to reliably recover, which is in line with the findings of Jaime Delgadillo et al. (2016). The effect size is similar to having a long-term health condition. Counter-intuitively, these patients are more likely to reliably improve and less likely to deteriorate. For service characteristics, patients in larger services are slightly less likely to reliably recover or improve and more likely to deteriorate. Unsurprisingly, patients in services with higher funding are more likely to reliably recover or improve and less likely to deteriorate.

In sum, the categories of patients that typically have lower mental health outcomes, e.g. living with a disability, also benefit less from the programme. Area deprivation is related negatively to patient outcomes, while funding of the services is positively related.

Results with data-driven sources of heterogeneity. Figure 3 presents the histograms of our heterogeneous treatment effect estimates produced by the generalised random forest described in Section 4.2.2.[38] The vertical dashed line again

[38]The forest includes 1,000 trees. Each tree is built using 10% of the sample. The minimum bin size is 500 observations. To improve the performance of the algorithm, some smaller covariate groups were merged together.

represents the estimated average treatment effect. The algorithm identifies some heterogeneity in treatment effects for all three outcomes. As in the previous approach, the distributions of treatment effects for reliable recovery and reliable improvement are bounded away from zero.

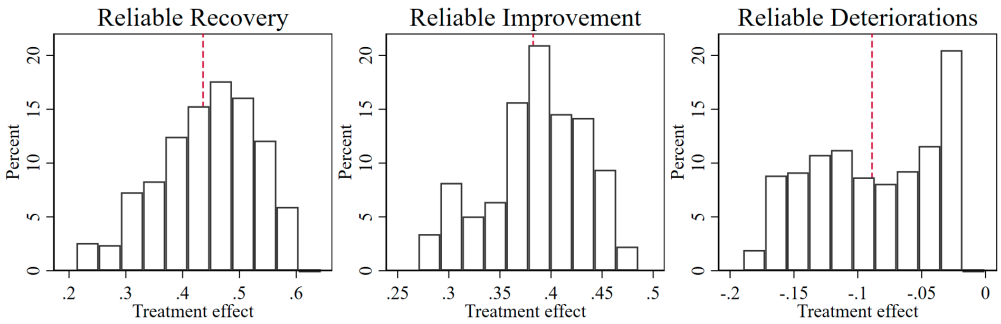


Figure 3: Conditional Average Treatment Effects – Generalised Random Forest

Note: The histograms plot the distributions of conditional average treatment effects estimated with generalised random forest.

To understand which sup-populations benefit most and least from treatment, we study the average levels of psychological-therapy and individual as well as service and local-area characteristics in sup-populations formed by quartiles of the estimated treatment effect distribution. The 1st quartile includes individuals whose estimated treatment effects were in the bottom 25% of all estimated individual treatment effects, quartiles 2 to 4 are formed accordingly. Appendix Tables E.III, E.IV, and E.V report the results for all covariates. Here, we discuss covariates that show substantial difference across quartiles.

Patients who recover less are more likely to be unemployed or long-term sick, or to have their gender, ethnicity, sexual orientation, or disability status not recorded. They are more likely to exhibit more severe symptoms at the start of treatment and less likely to self-refer. They are also more likely to live in deprived areas and attend larger services, as estimated by the number of patients. These patterns largely hold for reliable improvement, where, in addition, patients who are less likely to improve attend services which, on average, have lower funding.

For reliable deterioration, patterns are less clear. What is clear, however, is that patients for whom the programme is less effective, in terms of reducing deterioration, are more likely to experience more severe symptoms at the start of the treatment and to live in more deprived areas. We see less variation in other characteristics, including data quality (some characteristics not being recorded).

5.4 Cost-Benefit Calculations

We perform a simple and conservative back-of-the-envelope cost-benefit calculation of being treated within the IAPT programme. We appraise benefits and costs over a three-year period. Looking at benefits first, we found that treatment significantly decreases PHQ-9 scores by about five points, on average (cf. Table C.IV). A five-point decrease in PHQ-9 scores, in turn, corresponds to an increase in the *EuroQol-5 Dimensions (EQ-5D)* index of about 0.03 points (Furukawa et al., 2021).[39] UK Government values 1.0 QALYs at £70,000 (Treasury, 2022). For simplicity, let us assume that benefits accrue linearly over the course of treatment, which typically takes two months (corresponding to, on average, eight sessions, with one session per week). Unfortunately, the IAPT data do not include a long-run follow-up, so we cannot say something about relapse rates. However, the literature suggests that relapse rates after CBT are generally quite low (compared to alternative forms of treatment), typically around 40% six years after the end of treatment (cf. Fava et al., 2004). To be conservative, let us assume that relapse is instantaneous. With these considerations in mind, we obtain monetised benefits of =(((0.00 + 0.03) / 2) * 2 months + (0.03 * 0.6) * 10 months) / 12 months + 0.03 * 0.6 * 2 years * £70,000 = £3,745 per patient over a three-year period. Next, we look at costs. Clark (2018) calculates fixed costs per patient of £680 if one divides the total investment into IAPT in 2015–2016 (the start of our observation period, after which the programme reached its stable 50% target recovery rate) by the total number of courses of treatment during that period. Hence, we obtain net benefits of £3,745 - £680 = £3,065 per patient three years after the end of treatment, or a benefit-cost ratio of 5.5.[40]

This is likely to be a conservative ratio, for several reasons. When it comes to benefits, it is unlikely that relapse is instantaneous (in fact, Fava et al. (2004) show that relapse in the first twelve months after treatment is only about 15%). Moreover, we only looked at mental health, our main outcome. It is well-documented that improvements in mental health can lead to improvements in physical health later on (cf. Cho et al., 2010). We did not include ripple effects either, for exam-

[39] The EQ-5D is a routine instrument for the economic valuation of health-related quality of life, and its index is equivalent to a *Quality-Adjusted Life-Year (QALY)*, defined as one year in perfect mental and physical health. The index typically ranges from zero (representing death or a state equivalent to death, the worst possible health state) to one (representing full health, the best possible state). For more information on the instrument, see `https://euroqol.org/`.

[40] Using a discount rate of 3.5%, we obtain a benefit-cost ratio of 4.3.

ple on employment and productivity, nor spillovers on significant others (such as partners, children, or the wider community). For example, Smith et al. (2024) find that occupational income increases significantly two to three years after the end of psychological therapy. Reichman, Corman, and Noonan (2015) show that being out depression can lead to significant improvements in relationships. It is likely that these additional benefits are substantial. Most importantly, when it comes to costs, we only included direct programme costs, neglecting public savings to the exchequer in form of additional tax income and reduced (disability) benefits, nor did we include other savings to the healthcare system, which for the physically ill with co-morbid mental ill health can be substantial (Chiles, Lambert, and Hatch, 1999; Clark and Richard Layard, 2014). In fact, some authors argue that public savings in terms of taxes and benefits alone would turn net public costs negative, making the programme pay for itself (R. Layard, 2016).

6 Discussion and Conclusion

Mental ill health has profound impacts on individuals, their families, and society at large. It also remains a substantial challenge for the economy. Despite this, mental ill health is often relegated to the sidelines of healthcare priorities worldwide, overshadowed by physical health. This situation does not have to remain this way. There are now successful programmes that address mental ill health. When launched in England in 2008, the IAPT programme was the first, nationwide mental health service to make evidence-based psychological therapies for treating common mental health problems, in particular depression and anxiety, widely available to the general public. Still the largest in the world, the programme is regarded as a role model and is now being replicated in other countries.

This paper is the first to evaluate the casual impacts of this nationwide service at a scale that well represents the English population. Using data on all patients who started treatments between April 2016 and December 2018 and exploiting oversubscription and resulting exogenous variations in waiting times, we found that the programme provides significant mental health benefits. The mental health of treated patients' is more likely to have *reliably improved*, relative to a quasi-experimental waitlist control group, with a *reliable recovery* rate from mental ill health of about 43%. When exploring treatment heterogeneities, we found that, although the programme benefits all categories of patients we looked at, some

groups benefit less than others, e.g. those living with a disability, those residing in deprived areas, or those who were offered treatment not compliant with official guidelines.

We also found evidence of positive short-term effects of treatment beyond mental health outcomes. In particular, treated patients report less impairment in their work and social life due to mental ill health. Amongst those who were initially unemployed or on long-term sick leave, treated patients are more likely to report being employed and less likely to receive statutory sick pay at the end of treatment. Although these impacts are small, it should be noted that more sizeable labour market effects of psychological therapy have been found to materialise only two to three years after the end of treatment (cf. Smith et al., 2024). Taken together, being treated within the IAPT programme significantly and strongly improves patients' lives.

Our causal estimates of the IAPT treatment's effectiveness generally align qualitatively with previous findings from non-causal studies, which also observed improvements in patients after receiving treatment. However, the magnitudes of our estimates are smaller. The reason for this difference is that our quasi-experimental approach is able to isolate the treatment effect from natural recovery that happens over time.

Our back-of-the-envelope cost-benefit calculation of being treated within the IAPT programme shows that for every pound spent, the programme generates a benefit worth £5.50. This is likely to be a conservative estimate, as it does not account for ripple effects on physical health, employment and productivity, as well as spillovers on family members or the wider community. This estimate also overlooks potential future public savings in the form of additional tax income, reduced disability benefits, or savings to the healthcare system.

Our findings show that a nationwide mental health service "works" in providing evidence-based psychological therapies to the general public in a cost-effective manner. However, our work has limitations, some of which offer promising opportunities for future research. A notable extension of our analysis would involve evaluating the long-term impacts of the programme by collecting data that extend beyond the end of therapy, when systematic patient-level outcome monitoring stops. This prospective analysis would align closely with the ethos of the IAPT programme, which, from its start, has adopted a scientific evaluation mindset and can serve as a blueprint for the development of other public policies.

References

Angelucci, M. and D. Bennett (2023). "The Economic Impact of Depression Treatment in India: Evidence from Community-Based Provision of Pharmacotherapy". *American Economic Review* forthcoming.

Arias, D., S. Saxena, and S. Verguet (2022). "Quantifying the global burden of mental disorders and their economic value". *eClinicalMedicine: Part of The Lancet Discovery Science* 54, p. 101675.

Athey, Susan, Julie Tibshirani, and Stefan Wager (2019). "Generalized random forests". EN. *The Annals of Statistics* 47.2, pp. 1148–1178.

Avendano, M., A. de Coulon, and V. Nafilyan (2020). "Does longer compulsory schooling affect mental health? Evidence from a British reform". *Journal of Public Economics* 183, p. 104137.

Baranov, V. et al. (2020). "Maternal Depression, Women's Empowerment, and Parental Investment: Evidence from a Randomized Controlled Trial". *American Economic Review* 110.3, pp. 824–859.

Barker, N. et al. (2022). "Cognitive Behavioral Therapy among Ghana's Rural Poor Is Effective Regardless of Baseline Mental Distress". *American Economic Review: Insights* 4.4, pp. 527–545.

Barnay, T. and S. Juin (2016). "Does home care for dependent elderly people improve their mental health?" *Journal of Health Economics* 45, pp. 149–160.

Beam, E. A. and S. Quimbo (2023). "The Impact of Short-Term Employment for Low-Income Youth: Experimental Evidence from the Philippines". *Review of Economics and Statistics* 105.6, pp. 1379–1393.

Beck, J. S. (2020). *Cognitive Behavior Therapy: Basics and Beyond*. New York: Guilford Press.

Berger, M. C. and D. A. Black (1992). "Child Care Subsidies, Quality of Care, and the Labor Supply of Low-Income, Single Mothers". *Review of Economics and Statistics* 74.4, pp. 635–642.

Blattman, C., J. C. Jamison, and M. Sheridan (2017). "Reducing Crime and Violence: Experimental Evidence from Cognitive Behavioral Therapy in Liberia". *American Economic Review* 107.4, pp. 1165–1206.

Bossuroy, T. et al. (2022). "Tackling psychosocial and capital constraints to alleviate poverty". *Nature* 605, pp. 291–297.

Brewin, C. R. (1996). "Theoretical Foundations of Cognitive-Behavior Therapy for Anxiety and Depression". *Annual Review of Psychology* 47, pp. 33–57.

Cano-Vindel, Antonio et al. (2022). "Improving Access to Psychological Therapies in Spain: From IAPT to PsicAP". en. *Psicothema* 34.1, pp. 18–24.

Chernozhukov, Victor et al. (2018). "Double/debiased machine learning for treatment and structural parameters". en. *The Econometrics Journal* 21.1, pp. C1–C68.

Chiles, J. A., M. J. Lambert, and A. L. Hatch (1999). "The Impact of Psychological Interventions on Medical Cost Offset: A Meta-analytic Review". *Clinical Psychology: Science and Practice* 6.2, pp. 204–220.

Cho, H. J. et al. (2010). "Prior Depression History and Deterioration of Physical Health in Community-Dwelling Older Adults – A Prospective Cohort Study". *American Journal of Geriatric Psychiatry* 18.5, pp. 442–451.

Chuard, C. (2023). "Negative effects of long parental leave on maternal health: Evidence from a substantial policy change in Austria". *Journal of Health Economics* 88, p. 102726.

Clark, David M. (2018). "Realizing the Mass Public Benefit of Evidence-Based Psychological Therapies: The IAPT Program". eng. *Annual Review of Clinical Psychology* 14, pp. 159–183.

— (2019). *IAPT at 10: Achievements and challenges*. https://www.england.nhs.uk/blog/iapt-at-10-achievements-and-challenges/.

Clark, David M., L. Canvin, et al. (2018). "Transparency about the outcomes of mental health services (IAPT approach): an analysis of public data". *Lancet* 391, pp. 679–686.

Clark, David M., R. Layard, et al. (2009). "Improving access to psychological therapy: Initial evaluation of two UK demonstration sites". *Behaviour Research and Therapy* 47, pp. 910–920.

Clark, David M. and Richard Layard (2014). *Thrive: The Power of Psychological Therapy*. London: Penguin.

Clark, David M., J. Wild, et al. (2022). "More than doubling the clinical benefit of each hour of therapist time: a randomised controlled trial of internet cognitive therapy for social anxiety disorder". *Psychological Medicine* 53.11, pp. 5022–5032.

Connor, K. M. et al. (2000). "Psychometric properties of the Social Phobia Inventory (SPIN)". *British Journal of Psychiatry* 176.4, pp. 379–386.

Cuijpers, P., I. Cristea, et al. (2016). "How effective are cognitive behavior therapies for major depression and anxiety disorders? A meta-analytic update of the evidence". *World Psychiatry* 15.3, pp. 245–258.

Cuijpers, P., F. Smit, et al. (2010). "Efficacy of cognitive–behavioural therapy and other psychological treatments for adult depression: meta-analytic study of publication bias". *British Journal of Psychiatry* 196.3, pp. 173–178.

Dague, L., T. DeLeire, and L. Leininger (2017). "The Effect of Public Insurance Coverage for Childless Adults on Labor Supply". *American Economic Journal: Economic Policy* 9.2, pp. 124–154.

Delgadillo, J. et al. (2018). "On poverty, politics and psychology: the socioeconomic gradient of mental healthcare utilisation and outcomes". *British Journal of Psychiatry* 209.5, pp. 429–430.

Delgadillo, Jaime et al. (2016). "On poverty, politics and psychology: the socioeconomic gradient of mental healthcare utilisation and outcomes". en. *British Journal of Psychiatry* 209.5, pp. 429–430.

Department for Health (2008). *Speech by the Rt Hon Alan Johnson MP, Secretary of State for Health, 27 November 2008 at the New Savoy Partnership Annual Conference: Psychological therapies in the NHS: science, practice and policy*.

Digital, NHS (2016). *Psychological Therapies: Annual Report on the Use of IAPT Services - England, 2015-16*. Leeds: Health and Social Care Information Centre.

Dinerstein, M., R. Megalokonomou, and C. Yannelis (2022). "Human Capital Depreciation and Returns to Experience". *American Economic Review* 112.11, pp. 3725–3762.

Ehlers, A. et al. (2023). "Therapist-assisted online psychological therapies differing in trauma focus for post-traumatic stress disorder (STOP-PTSD): a UK-based, single-blind, randomised controlled trial". *Lancet Psychiatry* 10.8, pp. 608–622.

Fava, G. A. et al. (2004). "Six-Year Outcome of Cognitive Behavior Therapy for Prevention of Recurrent Depression". *American Journal of Psychiatry* 161.10, pp. 1872–1876.

Finkelstein, A., N. Hendren, and E. F. P. Luttmer (2019). "The Value of Medicaid: Interpreting Results from the Oregon Health Insurance Experiment". *Journal of Political Economy* 127.6, pp. 2836–2874.

Fonagy, P. et al. (2019). "Dynamic interpersonal therapy for moderate to severe depression: a pilot randomized controlled and feasibility trial ". *Psychological Medicine* 50.6, pp. 1010–1019.

Frisch, Ragnar and Frederick V. Waugh (1933). "Partial Time Regressions as Compared with Individual Trends". *Econometrica* 1.4, pp. 387–401.

Fryer Jr., R. G. and L. F. Katz (2013). "Achieving Escape Velocity: Neighborhood and School Interventions to Reduce Persistent Inequality". *American Economic Review* 103.3, pp. 232–237.

Furukawa, T. A. et al. (2021). "How can we estimate QALYs based on PHQ-9 scores? Equipercentile linking analysis of PHQ-9 and EQ-5D". *BMJ Mental Health* 24, pp. 97–101.

Ghosal, S. et al. (2022). "Sex Workers, Stigma and Self-Image: Evidence from Kolkata Brothels". *Review of Economics and Statistics* 104.3, pp. 431–448.

Gine, Evarist and Joel Zinn (1990). "Bootstrapping General Empirical Measures". *The Annals of Probability* 18.2, pp. 851–869.

Gruber, J. et al. (2022). "The impact of mental health support for the chronically ill on hospital utilisation: Evidence from the UK". *Social Science & Medicine* 294, p. 114675.

Gyani, A. et al. (2013). "Enhancing recovery rates: Lessons from year one of IAPT". *Behaviour Research and Therapy* 51.9, pp. 597–606.

Harvey, L. J. et al. (2023). "Investigating the efficacy of a Dialectical behaviour therapy-based universal intervention on adolescent social and emotional well-being outcomes". *Behaviour Research and Therapy* 169, p. 104408.

Haushofer, J., R. Mudida, and J. Shapiro (2022). *The Comparative Impact of Cash Transfers and a Psychotherapy Program on Psychological and Economic Well-being*. mimeo.

Heller, S. B. et al. (2017). "Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago". *Quarterly Journal of Economics* 132.1, pp. 1–54.

Hoe, T. P. (2023). "Does Hospital Crowding Matter? Evidence from Trauma and Orthopedics in England". *American Economic Journal: Economic Policy* 14.2, pp. 231–236.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press.

Imbens, G. W. and J. M. Wooldridge (2009). "Recent Developments in the Econometrics of Program Evaluation". *Journal of Economic Literature* 47.1, pp. 5–86.

Jacob, B. A., M. Kapustin, and J. Ludwig (2015). "The Impact of Housing Assistance on Child Outcomes: Evidence from a Randomized Housing Lottery". *Quarterly Journal of Economics* 130.1, pp. 465–506.

Jacob, B. A. and J. Ludwig (2012). "The Effects of Housing Assistance on Labor Supply: Evidence from a Voucher Lottery". *American Economic Review* 102.1, pp. 272–304.

Johnsen, T. J. and O. Friborg (2015). "The effects of cognitive behavioral therapy as an anti-depressive treatment is falling: A meta-analysis". *Psychological Bulletin* 141.4, pp. 747–768.

Kalin, N. H. (2020). "The Critical Relationship Between Anxiety and Depression". *American Journal of Psychiatry* 177.5, pp. 365–367.

Knapstad, Marit et al. (2020). "Effectiveness of Prompt Mental Health Care, the Norwegian Version of Improving Access to Psychological Therapies: A Randomized Controlled Trial". en. *Psychotherapy and Psychosomatics* 89.2, pp. 90–105.

Kroenke, K., R. L. Spitzer, and J. B. W. Williams (2001). "The PHQ-9: Validity of a Brief Depression Severity Measure". *Journal of General Internal Medicine* 16.9, pp. 606–613.

Lambert, M. J. (2013). *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change*. New York: Wiley.

Lang, M. (2013). "The Impact of Mental Health Insurance Laws on State Suicide Rates". *Health Economics* 22.1, pp. 73–88.

Layard, R. (2016). "The economics of mental health". *IZA World of Labor* 321, pp. 1–10.

Layard, R. and David M. Clark (2014). *Thrive: How Better Mental Health Care Transforms Lives and Saves Money*. Princeton, NJ: Princeton University Press.

List, J. A. (2022). *The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale*. New York: Penguin Random House.

Ludwig, J. et al. (2013). "Long-Term Neighborhood Effects on Low-Income Families: Evidence from Moving to Opportunity". *American Economic Review* 103.3, pp. 226–231.

Marcus, S. C. and M. Olfson (2010). "National Trends in the Treatment for Depression From 1998 to 2007". *Archives of General Psychiatry* 67.12, pp. 1265–1273.

McHugh, R. J. et al. (2013). "Patient Preference for Psychological vs Pharmacologic Treatment of Psychiatric Disorders: A Meta-Analytic Review". *Journal of Clinical Psychiatry* 74.6, pp. 595–602.

McManus, S., P. Bebbington, and R. Jenkins (2016). *Mental Health and Wellbeing in England: Adult Psychiatric Morbidity Survey 2014*. Leeds: NHS Digital.

Moller, N. P. et al. (2019). "The 2018 UK NHS Digital annual report on the Improving Access to Psychological Therapies programme: a brief commentary". *BMC Psychiatry* 19, p. 252.

Mundt, J. C. et al. (2002). "The Work and Social Adjustment Scale: a simple measure of impairment in functioning". *British Journal of Psychiatry* 180.5, pp. 461–464.

Nathan, P. E. and J. M. Gorman (2015). *A Guide to Treatments That Work*. Oxford: Oxford University Press.

NHS (2017). *Psychological Therapies, Annual Report on the Use of IAPT Services, England 2016-17*.

— (2019). *NHS Long-Term Plan*.

— (2021). *Psychological Therapies, Annual Report on the Use of IAPT Services, 2020-21*. https://digital.nhs.uk/data-and-information/publications/statistical/psychological-therapies-annual-reports-on-the-use-of-iapt-services/annual-report-2021-22.

NHS Confederation (2021). *What were clinical commissioning groups?*

NHS England (2019). *Mental Health Implementation Plan 2019/20 – 2023/24*.

Ortega, A. (2022). "Medicaid Expansion and mental health treatment: Evidence from the Affordable Care Act". *Health Economics* 32.4, pp. 755–806.

Reichman, N. E., H. Corman, and K. Noonan (2015). "Effects of maternal depression on couple relationship status". *Review of Economics of the Household* 13, pp. 929–973.

Richards, D., A. Enrique, et al. (2020). "A pragmatic randomized waitlist-controlled effectiveness and cost-effectiveness trial of digital interventions for depression and anxiety". *npj Digital Medicine* 3, p. 85.

Richards, D. and R. Suckling (2009). "Improving access to psychological therapies: Phase IV prospective cohort study". *British Journal of Clinical Psychology* 48.4, pp. 377–396.

Robles, S., M. Gross, and R. W. Fairlie (2021). "The effect of course shutouts on community college students: Evidence from waitlist cutoffs". *Journal of Public Economics* 199, p. 104409.

Roth, A. and D. Fonagy (2005). *What Works for Whom? A Critical Review of Psychotherapy Research*. New York: Guildford Press.

Roth, J. et al. (2023). "What's trending in difference-in-differences? A synthesis of the recent econometrics literature". *Journal of Econometrics* 235.2, pp. 2218–2244.

Rubin, D. B. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies". *Journal of Educational Psychology* 66.5, pp. 688–701.

Saunders, R., J. E. J. Buckman, and S. Pilling (2020). "Latent variable mixture modelling and individual treatment prediction". *Behavior Research and Therapy* 124, p. 103505.

Saunders, R., J. Cape, et al. (2016). "Predicting treatment outcome in psychological treatment services by identifying latent profiles of patients". *Journal of Affective Disorders* 197, pp. 107–115.

Sevim, D. et al. (2023a). *Socioemotional Skills in Early Childhood: Evidence from a Maternal Psychosocial Intervention*. IZA Discussion Paper 15925. Institute of Labor Economics.

— (2023b). *Trajectories of Early Childhood Skill Development and Maternal Mental Health*. Working Paper 1469. University of Warwick, Department of Economics.

Smith, O. R. F et al. (2024). *Cost-benefit of IAPT Norway and effects on work-related outcomes and health care utilization: results from a randomized controlled trial using registry-based data*. mimeo.

Spitzer, R. L. et al. (2006). "A brief measure for assessing generalized anxiety disorder: the GAD-7". *Archives of Internal Medicine* 166.10, pp. 1092–1097.

Stillman, S., D. McKenzie, and J. Gibson (2009). "Migration and mental health: Evidence from a natural experiment". *Journal of Health Economics* 28.3, pp. 677–687.

Strauss, C. et al. (2023). "Mindfulness-Based Cognitive Therapy Self-help Compared With Supported Cognitive Behavioral Therapy Self-help for Adults Experiencing Depression: The Low-Intensity Guided Help Through Mindfulness (LIGHTMind) Randomized Clinical Trial". *JAMA Psychiatry* 80.5, pp. 415–424.

Toffolutti, V. et al. (2021). "The employment and mental health impact of integrated Improving Access to Psychological Therapies: Evidence on secondary health care utilization from a pragmatic trial in three English counties". *Journal of Health Services Research Policy* 26.4, pp. 224–233.

Treasury, HM (2022). *The Green Book*. https://www.gov.uk/government/publications/the-green-book-appraisal-and-evaluation-in-central-government/the-green-book-2020.

Wager, Stefan and Susan Athey (2018). "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests". *Journal of the American Statistical Association* 113.523, pp. 1228–1242.

Wakefield, S. et al. (2020). "Improving Access to Psychological Therapies (IAPT) in the United Kingdom: A systematic review and meta-analysis of 10-years of practice-based evidence". *British Journal of Clinical Psychiatry* 60.1, pp. 1–37.

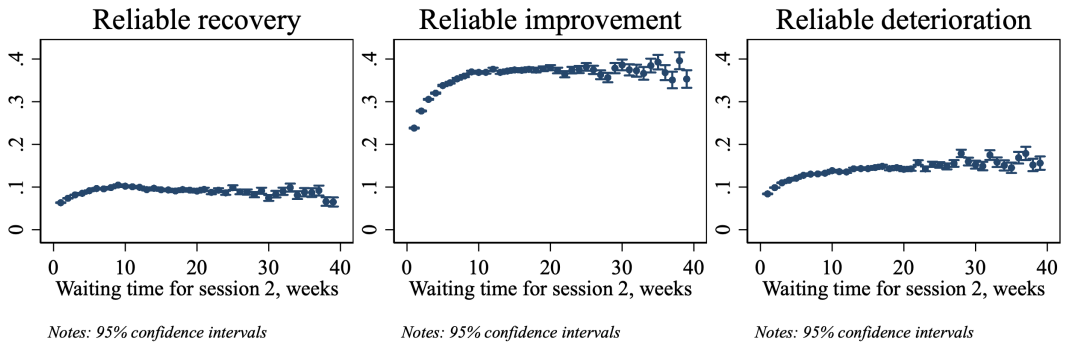
Williams, C. (2013). *Overcoming Anxiety, Stress and Panic: A Five Areas Approach*. Boca Raton, FL: CRC Press.

Williams, J. and A. L. Bretterville-Jensen (2022). *What's Another Day? The Effects of Wait Time for Substance Abuse Treatment on Health-Care Utilization, Employment and Crime*. IZA Discussion Paper 15083. IZA - Institute of Labor Economics.

Appendix

A Summary Statistics

Figure A.I: Main Outcomes for Different Waiting Times



Note: Own calculations.

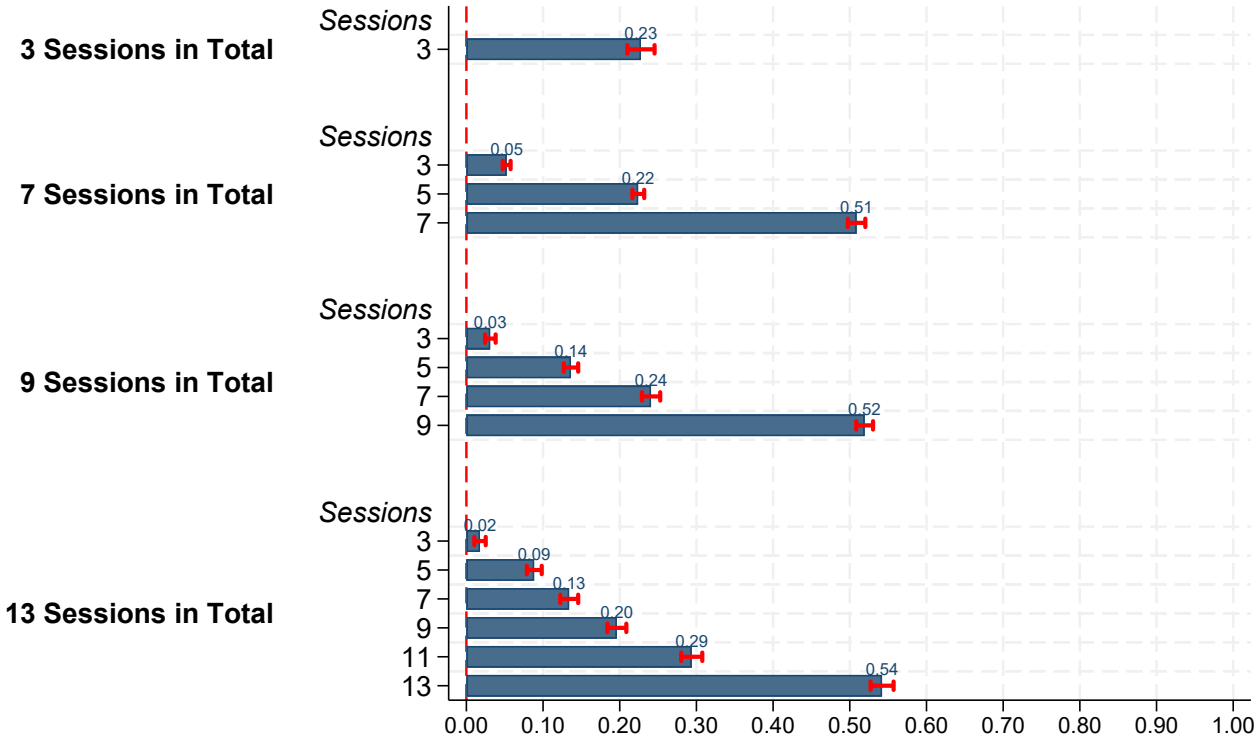


Figure A.II: Reliable Recovery – Session Value Added

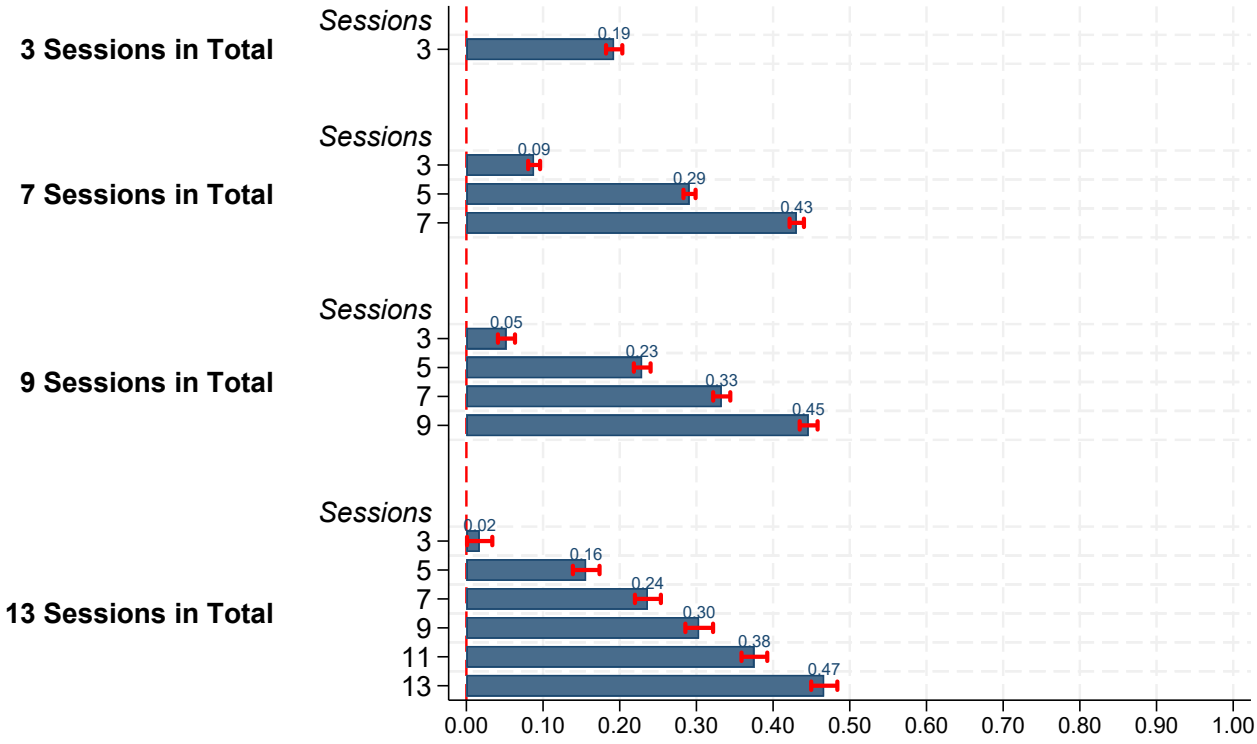


Figure A.III: Reliable Improvement – Session Value Added

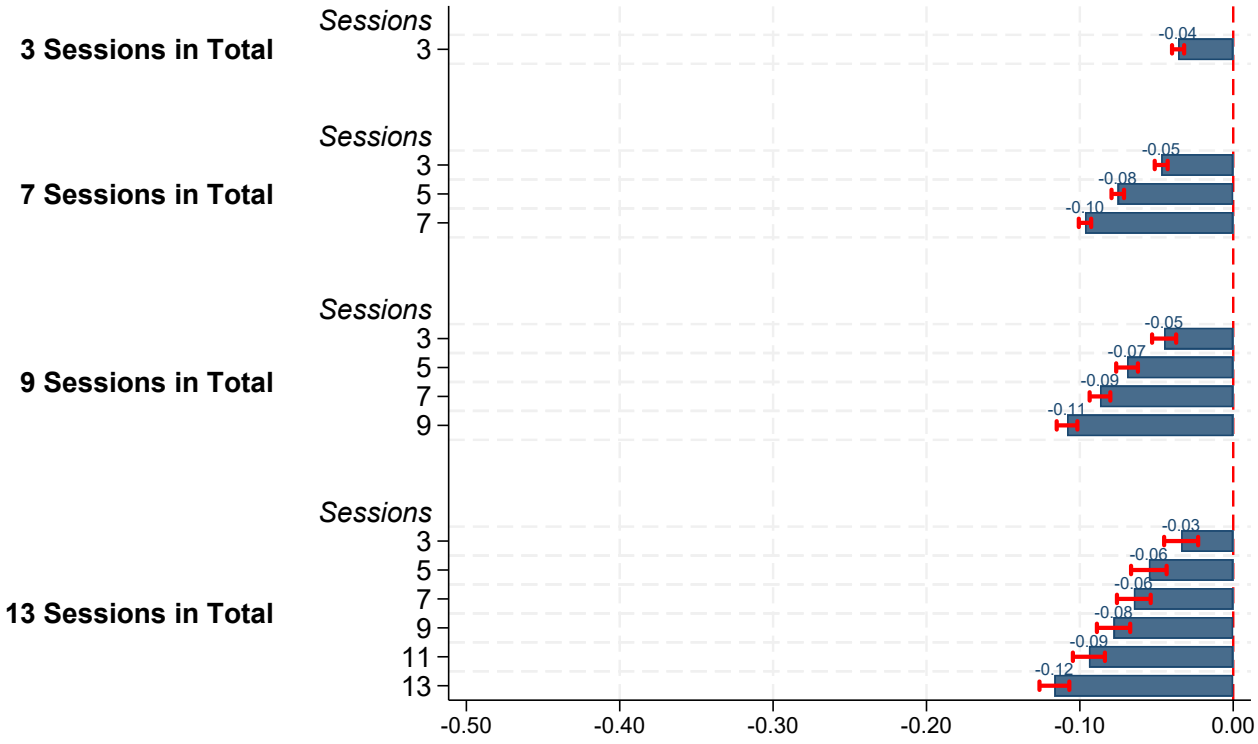


Figure A.IV: Reliable Deterioration – Session Value Added

Table A.I: Summary Statistics – Outcomes at Initial Assessment

	Average		Treatment Group		Control Group	
	Mean	SD	Mean	SD	Mean	SD
PHQ-9	15.765	5.497	15.688	5.504	15.841	5.490
GAD-7	14.389	4.338	14.310	4.350	14.468	4.324
Mental Health Index	0.434	0.685	0.421	0.686	0.446	0.683
Work and Social Adjustment Scale - Overall	20.044	9.229	19.849	9.153	20.236	9.298
Work and Social Adjustment Scale - Work	4.372	2.596	4.380	2.587	4.365	2.604
Work and Social Adjustment Scale - Home Management	3.620	2.393	3.584	2.369	3.656	2.416
Work and Social Adjustment Scale - Social Leisure	4.492	2.447	4.438	2.431	4.545	2.461
Work and Social Adjustment Scale - Private Leisure	3.687	2.541	3.634	2.515	3.739	2.564
Work and Social Adjustment Scale - Close Relationships	3.957	2.468	3.916	2.451	3.996	2.483
Employed (As Opposed To Unemployed)	0.857	0.350	0.858	0.349	0.856	0.351
Employed (As Opposed To Long-Term Sick)	0.880	0.324	0.894	0.308	0.867	0.339
Receiving Statutory Sick Pay	0.077	0.267	0.084	0.278	0.071	0.257

Table A.II: Summary Statistics – Covariates at Initial Assessment

Covariate	Mean	SD
Age	40.200	14.907
Gender: Male	0.247	0.432
Female	0.496	0.500
Non-Binary	0.000	0.022
No Response	0.256	0.436
Ethnicity: British	0.595	0.491
Irish	0.006	0.075
Any Other White Background	0.032	0.175
White and Black Caribbean	0.006	0.076
White and Black African	0.002	0.039
White and Asian	0.003	0.054
Any Other Mixed Background	0.006	0.077
Indian	0.014	0.116
Pakistani	0.010	0.099
Bangladeshi	0.003	0.056
Any Other Asian Background	0.007	0.086
Caribbean	0.010	0.098
African	0.007	0.085
Any Other Black Background	0.003	0.055
Chinese	0.002	0.041
Any Other Ethnic Group	0.009	0.094
No Response	0.287	0.452
Religion: Baha'i	0.000	0.010
Buddhist	0.002	0.050
Christian	0.190	0.393
Hindu	0.004	0.067
Jew	0.002	0.047
Muslim	0.020	0.139
Pagan	0.001	0.035
Sikh	0.004	0.060
Zoroastrian	0.000	0.008
Other	0.020	0.141
Not Religious	0.328	0.470
No Response	0.427	0.495
Sexual Orientation: Heterosexual or Straight	0.564	0.496
Gay or Lesbian	0.017	0.128
Bisexual	0.014	0.117
Other	0.009	0.094
No Response	0.397	0.489
Long-Term Health Condition: Yes	0.202	0.402
No	0.452	0.498
No Response	0.345	0.476
Employment Status: Employed	0.569	0.495
Unemployed and Seeking Work	0.095	0.293
Student	0.054	0.226
Long-Term Sick or Disabled	0.077	0.267
Homemaker Looking After a Family or Home	0.049	0.215

Not Receiving Benefits and Not Working	0.023	0.151
Unpaid Voluntary Work and Not Working or Actively Seeking	0.004	0.060
Retired	0.070	0.256
Refused	0.000	0.001
No Response	0.058	0.235
Services Member: Yes	0.000	0.015
Former	0.013	0.114
Not Former or Their Dependent	0.566	0.496
Dependent of Services Member	0.000	0.009
Dependent of Former Services Member	0.003	0.050
No Response	0.418	0.493
Mental Health Index	0.434	0.685
Referral: Acute Secondary Care	0.007	0.081
Child Health	0.000	0.016
Employer	0.000	0.022
IAPT Stepped Care	0.004	0.064
Independent/Voluntary Sector	0.004	0.062
Internal Referral	0.000	0.010
Internal Referral From Inpatient Service (Within Own NHS Trust)	0.000	0.009
Internal Referral from Community Mental Health Team	0.018	0.134
Justice System	0.001	0.031
Local Authority Services	0.001	0.033
Other	0.029	0.168
Other Mental Health NHS Trust	0.000	0.018
Primary Health Care	0.217	0.412
Self-Referral	0.715	0.451
Transfer by Graduation (Within Own NHS Trust)	0.000	0.009
Unknown	0.000	0.001
Referral Time Lapsed	3.029	3.713
Treatment Mode: Face-to-Face Communication	0.279	0.449
Telephone	0.684	0.465
Telemedicine	0.009	0.096
Talk Type for Person Unable to Speak	0.000	0.009
E-Mail	0.017	0.128
Text Messaging	0.002	0.040
Online Triage	0.000	0.004
No Response	0.008	0.092
Medication: Prescribed But Not Taking	0.045	0.208
Prescribed and Taking	0.477	0.499
Not Prescribed	0.415	0.493
No Response	0.063	0.243
Initial Diagnosis: Agoraphobia	0.007	0.083
Generalised Anxiety Disorder	0.221	0.415
Mixed Anxiety and Depressive Disorder	0.111	0.314
Obsessive-Compulsive Disorder	0.023	0.149
Other Anxiety or Stress-Related Disorder	0.039	0.193
Panic Disorder (Episodic Paroxysmal Anxiety)	0.028	0.166
Post-Traumatic Stress Disorder	0.041	0.198
Social Phobias	0.028	0.165
Specific (Isolated) Phobias	0.008	0.087

Depression	0.373	0.484
Invalid Data Supplied	0.001	0.031
Other Mental Health Problem	0.043	0.204
Other Recorded Problem	0.012	0.109
No Response	0.065	0.247
Treatment Intensity: Low Intensity	0.395	0.489
High Intensity	0.221	0.415
Step Up: Low to High Intensity	0.036	0.185
Step Down: High to Low Intensity	0.311	0.463
Multiple Changes in Intensity	0.037	0.189
CCG Number of Staff	116.387	90.115
CCG Number of Registered Patients	31,231.043	18,634.715
CCG Allocations Per Registered Patient	1,272.071	205.494
CCG Unemployment Rate	4.367	1.302
CCG Median Wage	457.250	69.245
Index of Multiple Deprivation: Average Rank	97.626	56.962
Income: Average Rank	16,810.156	4,453.149
Employment: Average Rank	16,724.635	4,657.311
Education, Skills, and Training: Average Rank	16,585.929	4,236.536
Health Deprivation and Disability: Average Rank	16,819.675	6,320.952
Crime: Average Rank	16,882.870	5,232.891
Barriers to Housing and Services: Average Rank	16,596.357	5,466.127
Living Environment: Average Rank	16,756.243	6,099.622

Table A.III: Balancing Properties of Covariates Between Treatment and Default Control Group (50th Percentile of Waiting Time)

	Treatment		Control		Overlap Measures			
	$N_T = 618,574$		$N_c = 628,218$		Norm.	Log Ratio	$\pi^{0.05}$	
	Mean	SD	Mean	SD	Diff.	of STD	Treatment	Control
Initial Assessment								
Reliable Recovery	0.000	0.000	0.000	0.000	0.000	-	0.000	0.000
Reliable Improvement	0.000	0.000	0.000	0.000	0.000	-	0.000	0.000
Reliable Deterioration	0.000	0.000	0.000	0.000	0.000	-	0.000	0.000
PHQ-9	15.688	5.504	15.841	5.490	-0.028	0.002	0.048	0.048
GAD-7	14.310	4.350	14.468	4.324	-0.037	0.006	0.017	0.016
Mental Health Index	0.421	0.686	0.446	0.683	-0.037	0.005	0.050	0.047
WSAS - Overall	19.849	9.153	20.236	9.298	-0.042	-0.016	0.115	0.116
WSAS - Work	4.380	2.587	4.365	2.604	0.006	-0.006	0.416	0.408
WSAS - Home Management	3.584	2.369	3.656	2.416	-0.030	-0.020	0.075	0.066
WSAS - Social Leisure	4.438	2.431	4.545	2.461	-0.044	-0.012	0.075	0.067
WSAS - Private Leisure	3.634	2.515	3.739	2.564	-0.041	-0.019	0.075	0.066
WSAS - Close Relationships	3.916	2.451	3.996	2.483	-0.032	-0.013	0.075	0.066
Employed (Not Unemployed)	0.858	0.349	0.856	0.351	0.007	-0.007	0.499	0.511
Employed (Not Long-Term Sick)	0.894	0.308	0.867	0.339	0.081	-0.096	0.562	0.532
Receiving Statutory Sick Pay	0.084	0.278	0.071	0.257	0.050	0.080	0.093	0.069
First Clinical Session								
Reliable Recovery	0.068	0.252	0.092	0.289	-0.088	-0.137	0.000	0.000
Reliable Improvement	0.273	0.445	0.356	0.479	-0.180	-0.072	0.000	0.000
Reliable Deterioration	0.105	0.306	0.136	0.342	-0.096	-0.113	0.000	0.000
PHQ-9	14.380	5.943	14.115	6.066	0.044	-0.020	0.038	0.043
GAD-7	13.180	4.942	12.972	5.111	0.041	-0.034	0.015	0.021
Mental Health Index	0.225	0.793	0.187	0.820	0.047	-0.033	0.037	0.056
WSAS - Overall	18.872	9.269	18.488	9.434	0.041	-0.018	0.116	0.126
WSAS - Work	4.065	2.590	3.796	2.572	0.104	0.007	0.435	0.405
WSAS - Home Management	3.476	2.298	3.442	2.342	0.015	-0.019	0.092	0.070
WSAS - Social Leisure	4.228	2.399	4.205	2.451	0.009	-0.022	0.092	0.070
WSAS - Private Leisure	3.489	2.419	3.401	2.461	0.036	-0.017	0.092	0.070
WSAS - Close Relationships	3.686	2.380	3.647	2.404	0.016	-0.010	0.092	0.070
Employed (Not Unemployed)	0.860	0.347	0.860	0.347	-0.001	0.001	0.554	0.561
Employed (Not Long-Term Sick)	0.893	0.309	0.863	0.344	0.093	-0.108	0.614	0.566
Receiving Statutory Sick Pay	0.074	0.261	0.048	0.214	0.107	0.199	0.136	0.103
Last Clinical Session								
Reliable Recovery	0.536	0.499	0.525	0.499	0.022	-0.001	0.000	0.000
Reliable Improvement	0.745	0.436	0.742	0.438	0.007	-0.004	0.000	0.000
Reliable Deterioration	0.050	0.219	0.057	0.231	-0.028	-0.056	0.000	0.000
PHQ-9	8.737	6.454	8.957	6.552	-0.034	-0.015	0.018	0.020
GAD-7	7.879	5.616	8.072	5.704	-0.034	-0.016	0.000	0.000
Mental Health Index	-0.657	0.935	-0.623	0.950	-0.035	-0.016	0.049	0.053
WSAS - Overall	12.622	9.815	12.883	9.986	-0.026	-0.017	0.105	0.094
WSAS - Work	2.742	2.472	2.702	2.471	0.016	0.001	0.443	0.432
WSAS - Home Management	2.405	2.163	2.474	2.206	-0.032	-0.020	0.085	0.096
WSAS - Social Leisure	2.757	2.352	2.829	2.394	-0.031	-0.018	0.085	0.070
WSAS - Private Leisure	2.260	2.220	2.331	2.265	-0.032	-0.020	0.085	0.070
WSAS - Close Relationships	2.477	2.249	2.521	2.270	-0.019	-0.010	0.085	0.070
Employed (Not Unemployed)	0.866	0.341	0.864	0.342	0.003	-0.004	0.547	0.560
Employed (Not Long-Term Sick)	0.888	0.315	0.860	0.347	0.084	-0.095	0.587	0.553

Receiving Statutory Sick Pay	0.040	0.197	0.030	0.172	0.054	0.138	0.118	0.097

Note: WSAS: Working and Social Adjustment Scale (Mundt et al., 2002). The normalised difference is calculated as $\Delta x = (\bar{x_t} - \bar{x_c})/\sqrt{(\sigma_t^2 + \sigma_c^2)}$, where $\bar{x_t}$ and $\bar{x_c}$ is the sample mean of variable x in the treatment and control group, respectively. σ^2 denotes the respective variance. A normalised difference greater than 0.25 indicates unbalancedness. The log of the ratio of standard deviations is calculated as $LR = ln(\frac{\sigma_t}{\sigma_c})$. The share of the control (treated) units outside the 0.025 and 0.975 quantiles of the covariate distribution of the treated (control) units is calculate as $(1 - F_t(F_c^{-1}(1 - \alpha/2))) + F_t(F_c^{-1}(\alpha/2))$ for treatment and $(1 - F_c(F_t^{-1}(1 - \alpha/2))) + F_c(F_t^{-1}(\alpha/2))$ (Imbens and Wooldridge, 2009; Imbens and Rubin, 2015).

Table A.IV: Balancing Properties of Covariates Between Treatment and Default Control Group (50th Percentile of Waiting Time)

	Treatment $N_T = 618,574$		Control $N_c = 628,218$		Overlap Measures			
	Mean	SD	Mean	SD	Norm. Diff.	Log Ratio of STD	$\pi^{0.05}$ Treatment	Control
Age	39.975	14.924	40.420	14.887	-0.030	0.002	0.042	0.041
Gender: Male	0.247	0.431	0.248	0.432	-0.003	-0.002	0.000	0.000
Female	0.489	0.500	0.504	0.500	-0.031	0.000	0.000	0.000
Non-Binary	0.000	0.022	0.000	0.022	-0.001	-0.022	0.000	0.000
No Response	0.264	0.441	0.247	0.432	0.038	0.021	0.000	0.000
Ethnicity: British	0.594	0.491	0.596	0.491	-0.005	0.001	0.000	0.000
Irish	0.005	0.073	0.006	0.077	-0.007	-0.045	0.000	0.000
Any Other White Background	0.030	0.171	0.033	0.179	-0.017	-0.045	0.000	0.000
White and Black Caribbean	0.005	0.074	0.006	0.078	-0.008	-0.053	0.000	0.000
White and Black African	0.001	0.038	0.002	0.040	-0.005	-0.063	0.000	0.000
White and Asian	0.003	0.055	0.003	0.054	0.001	0.011	0.000	0.000
Any Other Mixed Background	0.005	0.074	0.006	0.080	-0.012	-0.079	0.000	0.000
Indian	0.013	0.112	0.014	0.119	-0.015	-0.063	0.000	0.000
Pakistani	0.009	0.094	0.011	0.104	-0.021	-0.103	0.000	0.000
Bangladeshi	0.002	0.048	0.004	0.062	-0.029	-0.262	0.000	0.000
Any Other Asian Background	0.007	0.083	0.008	0.089	-0.013	-0.073	0.000	0.000
Caribbean	0.009	0.096	0.010	0.100	-0.008	-0.039	0.000	0.000
African	0.007	0.081	0.008	0.088	-0.014	-0.084	0.000	0.000
Any Other Black Background	0.003	0.052	0.003	0.057	-0.009	-0.086	0.000	0.000
Chinese	0.002	0.040	0.002	0.042	-0.004	-0.043	0.000	0.000
Any Other Ethnic Group	0.008	0.090	0.010	0.099	-0.019	-0.098	0.000	0.000
No Response	0.296	0.457	0.278	0.448	0.041	0.019	0.000	0.000
Religion: Baha'i	0.000	0.010	0.000	0.009	0.001	0.034	0.000	0.000
Buddhist	0.003	0.051	0.002	0.048	0.005	0.051	0.000	0.000
Christian	0.184	0.388	0.197	0.398	-0.033	-0.026	0.000	0.000
Hindu	0.004	0.064	0.005	0.070	-0.012	-0.091	0.000	0.000
Jew	0.002	0.044	0.003	0.050	-0.012	-0.131	0.000	0.000
Muslim	0.017	0.128	0.023	0.150	-0.045	-0.156	0.000	0.000
Pagan	0.001	0.034	0.001	0.036	-0.003	-0.043	0.000	0.000
Sikh	0.003	0.056	0.004	0.064	-0.015	-0.124	0.000	0.000
Zoroastrian	0.000	0.008	0.000	0.007	0.003	0.222	0.000	0.000
Other	0.019	0.137	0.021	0.144	-0.015	-0.050	0.000	0.000
Not Religious	0.324	0.468	0.333	0.471	-0.019	-0.007	0.000	0.000
No Response	0.443	0.497	0.411	0.492	0.065	0.010	0.000	0.000
Sexual Orientation: Heterosexual or Straight	0.552	0.497	0.576	0.494	-0.049	0.006	0.000	0.000
Gay or Lesbian	0.016	0.126	0.017	0.130	-0.009	-0.033	0.000	0.000
Bisexual	0.014	0.116	0.014	0.118	-0.004	-0.017	0.000	0.000
Other	0.008	0.088	0.010	0.100	-0.023	-0.118	0.000	0.000
No Response	0.411	0.492	0.383	0.486	0.057	0.012	0.000	0.000
Long-Term Health Condition: Yes	0.196	0.397	0.208	0.406	-0.031	-0.023	0.000	0.000
No	0.452	0.498	0.453	0.498	-0.003	0.000	0.000	0.000
No Response	0.352	0.478	0.339	0.473	0.029	0.009	0.000	0.000
Employment Status: Employed	0.572	0.495	0.566	0.496	0.012	-0.002	0.000	0.000
Unemployed and Seeking Work	0.095	0.293	0.096	0.294	-0.003	-0.004	0.000	0.000
Student	0.055	0.228	0.053	0.224	0.009	0.017	0.000	0.000
Long-Term Sick or Disabled	0.068	0.252	0.087	0.281	-0.069	-0.111	0.000	0.000
Homemaker Looking After a Family or Home	0.049	0.215	0.048	0.215	0.002	0.003	0.000	0.000

Not Receiving Benefits and Not Working	0.021	0.145	0.025	0.157	-0.026	-0.082	0.000	0.000
Unpaid Voluntary Work and Not Working or Actively Seeking	0.003	0.059	0.004	0.060	-0.003	-0.023	0.000	0.000
Retired	0.069	0.254	0.071	0.257	-0.007	-0.012	0.000	0.000
Refused	0.000	0.000	0.000	0.001	-0.002	-	0.000	0.000
No Response	0.067	0.250	0.050	0.218	0.072	0.136	0.000	0.000
Services Member: Yes	0.000	0.020	0.000	0.007	0.024	1.091	0.000	0.000
Former	0.014	0.119	0.012	0.109	0.019	0.082	0.000	0.000
Not Former or Their Dependent	0.548	0.498	0.583	0.493	-0.072	0.009	0.000	0.000
Dependent of Services Member	0.000	0.008	0.000	0.010	-0.004	-0.222	0.000	0.000
Dependent of Former Services Member	0.002	0.050	0.003	0.050	-0.001	-0.009	0.000	0.000
No Response	0.435	0.496	0.402	0.490	0.067	0.011	0.000	0.000
Mental Health Index	0.421	0.686	0.446	0.683	-0.037	0.005	0.050	0.047
Referral: Acute Secondary Care	0.007	0.083	0.006	0.079	0.008	0.048	0.000	0.000
Child Health	0.000	0.016	0.000	0.016	-0.001	-0.037	0.000	0.000
Employer	0.001	0.025	0.000	0.018	0.013	0.301	0.000	0.000
IAPT Stepped Care	0.005	0.074	0.003	0.053	0.042	0.337	0.000	0.000
Independent/Voluntary Sector	0.004	0.066	0.003	0.057	0.020	0.161	0.000	0.000
Internal Referral	0.000	0.011	0.000	0.010	0.002	0.114	0.000	0.000
Internal Referral From Inpatient Service (Within Own NHS Trust)	0.000	0.009	0.000	0.009	0.000	0.027	0.000	0.000
Internal Referral from Community Mental Health Team	0.016	0.125	0.020	0.142	-0.034	-0.123	0.000	0.000
Justice System	0.001	0.037	0.001	0.025	0.023	0.377	0.000	0.000
Local Authority Services	0.001	0.035	0.001	0.030	0.009	0.133	0.000	0.000
Other	0.032	0.175	0.027	0.161	0.029	0.081	0.000	0.000
Other Mental Health NHS Trust	0.000	0.017	0.000	0.019	-0.003	-0.086	0.000	0.000
Primary Health Care	0.206	0.405	0.227	0.419	-0.050	-0.035	0.000	0.000
Self-Referral	0.721	0.448	0.709	0.454	0.028	-0.013	0.000	0.000
Transfer by Graduation (Within Own NHS Trust)	0.000	0.007	0.000	0.010	-0.005	-0.307	0.000	0.000
Unknown	0.000	0.000	0.000	0.001	-0.002	-	0.000	0.000
Referral Time Lapsed	3.227	4.370	2.833	2.912	0.106	0.406	0.049	0.009
Treatment Mode: Face-to-Face Communication	0.345	0.475	0.214	0.410	0.295	0.147	0.000	0.000
Telephone	0.606	0.489	0.761	0.426	-0.337	0.136	0.000	0.000
Telemedicine	0.018	0.133	0.001	0.028	0.179	1.552	0.000	0.000
Talk Type for Person Unable to Speak	0.000	0.011	0.000	0.007	0.008	0.434	0.000	0.000
E-Mail	0.020	0.140	0.013	0.115	0.053	0.202	0.000	0.000
Text Messaging	0.001	0.035	0.002	0.045	-0.018	-0.232	0.000	0.000
Online Triage	0.000	0.005	0.000	0.003	0.005	0.760	0.000	0.000
No Response	0.009	0.093	0.008	0.090	0.007	0.037	0.000	0.000
Medication: Prescribed But Not Taking	0.043	0.204	0.047	0.213	-0.019	-0.042	0.000	0.000
Prescribed and Taking	0.464	0.499	0.489	0.500	-0.051	-0.002	0.000	0.000
Not Prescribed	0.416	0.493	0.414	0.492	0.005	0.001	0.000	0.000
No Response	0.076	0.266	0.049	0.217	0.111	0.203	0.000	0.000
Initial Diagnosis: Agoraphobia	0.006	0.079	0.007	0.086	-0.014	-0.085	0.000	0.000
Generalised Anxiety Disorder	0.222	0.415	0.219	0.414	0.006	0.004	0.000	0.000
Mixed Anxiety and Depressive Disorder	0.119	0.324	0.103	0.304	0.051	0.064	0.000	0.000
Obsessive-Compulsive Disorder	0.021	0.143	0.025	0.155	-0.027	-0.085	0.000	0.000
Other Anxiety or Stress-Related Disorder	0.037	0.189	0.040	0.197	-0.017	-0.040	0.000	0.000

Panic Disorder (Episodic Paroxysmal Anxiety)	0.029	0.167	0.028	0.166	0.001	0.004	0.000	0.000
Post-Traumatic Stress Disorder	0.036	0.187	0.046	0.209	-0.048	-0.112	0.000	0.000
Social Phobias	0.026	0.158	0.030	0.171	-0.028	-0.080	0.000	0.000
Specific (Isolated) Phobias	0.007	0.084	0.008	0.090	-0.012	-0.071	0.000	0.000
Depression	0.362	0.481	0.384	0.486	-0.046	-0.012	0.000	0.000
Invalid Data Supplied	0.001	0.033	0.001	0.029	0.008	0.123	0.000	0.000
Other Mental Health Problem	0.047	0.212	0.040	0.195	0.036	0.080	0.000	0.000
Other Recorded Problem	0.011	0.107	0.013	0.112	-0.011	-0.051	0.000	0.000
No Response	0.076	0.265	0.055	0.228	0.084	0.149	0.000	0.000
Treatment Intensity: Low Intensity	0.397	0.489	0.392	0.488	0.009	0.002	0.000	0.000
High Intensity	0.220	0.415	0.222	0.416	-0.004	-0.003	0.000	0.000
Step Up: Low to High Intensity	0.035	0.184	0.036	0.186	-0.005	-0.012	0.000	0.000
Step Down: High to Low Intensity	0.310	0.463	0.312	0.463	-0.005	-0.002	0.000	0.000
Multiple Changes in Intensity	0.037	0.190	0.037	0.189	0.003	0.007	0.000	0.000
CCG Number of Staff	119.737	93.331	113.089	86.706	0.074	0.074	0.072	0.038
CCG Number of Registered Patients	31,551.943	18,936.964	30,915.069	18,326.762	0.034	0.033	0.054	0.041
CCG Allocations Per Registered Patient	1,259.523	225.230	1,284.427	183.167	-0.121	0.207	0.056	0.061
CCG Unemployment Rate	4.360	1.335	4.373	1.269	-0.010	0.051	0.058	0.043
CCG Median Wage	454.474	67.593	459.984	70.727	-0.080	-0.045	0.052	0.053
Index of Multiple Deprivation: Average Rank	99.195	57.403	96.083	56.482	0.055	0.016	0.054	0.044
Income: Average Rank	16,648.934	4,489.914	16,968.902	4,410.900	-0.072	0.018	0.050	0.051
Employment: Average Rank	16,616.696	4,701.454	16,830.916	4,610.969	-0.046	0.019	0.053	0.051
Education, Skills, and Training: Average Rank	16,650.542	4,187.294	16,522.309	4,283.521	0.030	-0.023	0.051	0.043
Health Deprivation and Disability: Average Rank	16,721.574	6,333.467	16,916.271	6,307.118	-0.031	0.004	0.051	0.053
Crime: Average Rank	16,739.634	5,245.765	17,023.908	5,216.346	-0.054	0.006	0.047	0.050
Barriers to Housing and Services: Average Rank	16,584.651	5,248.194	16,607.885	5,672.520	-0.004	-0.078	0.042	0.060
Living Environment: Average Rank	16,635.006	5,985.810	16,875.619	6,207.341	-0.039	-0.036	0.046	0.055

Note: The normalised difference is calculated as $\Delta x = (\bar{x_t} - \bar{x_c})/\sqrt{(\sigma_t^2 + \sigma_c^2)}$, where $\bar{x_t}$ and $\bar{x_c}$ is the sample mean of variable x in the treatment and control group, respectively. σ^2 denotes the respective variance. A normalised difference greater than 0.25 indicates unbalancedness. The log of the ratio of standard deviations is calculated as $LR = ln(\frac{\sigma_t}{\sigma_c})$. The share of the control (treated) units outside the 0.025 and 0.975 quantiles of the covariate distribution of the treated (control) units is calculate as $(1 - F_t(F_c^{-1}(1-\alpha/2))) + F_t(F_c^{-1}(\alpha/2))$ for treatment and $(1 - F_c(F_t^{-1}(1-\alpha/2))) + F_c(F_t^{-1}(\alpha/2))$ (Imbens and Wooldridge, 2009; Imbens and Rubin, 2015).

B Identification and Estimation Proofs

Proposition 1 proves that Assumptions 1 and 2 enable us to identify ATT and CATT.

Proposition 1. Under Assumptions 1 and 2, ATT and CATT are identified from the joint distribution of $(\Delta Y_{it_i}, D_{it_i}, X_{it_i})$.

Proof. Under Assumption 1, expanding out $\Delta Y_{it_i}(0)$ and re-arrange gives:

$$E\left[Y_{it_{i2}}(0)\,|D_{it_i}=1, X_{it_i}\right] = E\left[Y_{it_{i1}}(0)\,|D_{it_i}=1, X_{it_i}\right] + E\left[\Delta Y_{it_i}(0)\,|D_{it_i}=0, X_{it_i}\right].$$

By Assumption 2, the first term on the right-hand-side of the equation above becomes $E\left[Y_{it_{i1}}(1)\,|D_{it_i}=1, X_{it_i}\right]$, so that $E\left[Y_{it_{i2}}(0)\,|D_{it_i}=1, X_{it_i}\right]$ is equal to $E\left[Y_{it_{i1}}|D_{it_i}=1, X_{it_i}\right] + E\left[Y_{it_{i2}} - Y_{it_{i1}}|D_{it_i}=0, X_{it_i}\right]$. Subsequently, CATT is identified from the joint distribution of $(\Delta Y_{it_i}, D_{it_i}, X_{it_i})$ since,

$$\theta\left(X_{it_i}\right) = E\left[\Delta Y_{it_i}|D_{it_i}=1, X_{it_i}\right] - E\left[\Delta Y_{it_i}|D_{it_i}=0, X_{it_i}\right].$$

Subsequently, ATT is also identified because, by the law of iterated expectation, $\theta = E\left[\theta\left(X_{it_i}\right)|D_{it_i}=1\right]$.■

The proof strategy used in Proposition 1 is the conditional version of the was used in Section 2 of J. Roth et al., 2023. J. Roth et al., 2023 also discussed the importance of another condition for nonparametric inference known as *Strong Overlap* (see their Assumption 7), which requires $P\left(D_{it_i}|X_{it_i}\right)$ to be uniformly bounded away from 1 almost surely and $E\left[D_{it_i}\right] > 0$. The Strong Overlap condition is clearly supported empirically by our estimating sample as we have numerous untreated patients for every combination of covariates observed and we have a large shares of treated and untreated patients unconditionally.

Proposition 2 proves our nonparametric estimator for $\{\theta(w,q)\}$ can be obtained from OLS estimation.

Proposition 2. OLS estimator of $\theta(w,q)$ in equation (6) is the same as the nonparametric matching estimator in Section 4.2.2.

Proof. We start by re-writing equation (6) as,

$$\Delta Y_{it_i} = \sum_{w,q}\left[\beta(w,q) + \theta(w,q)\times D_{it_i}\right]\times \mathbf{1}\{Q_{it_i}=q, W_{it_i}=w\} + u_{it_i},$$

which has the following matrix representation,

$$\mathbf{\Delta Y} = \sum_{w,q}\left[\iota(w,q) : \mathbf{D}(w,q)\right]\begin{bmatrix}\beta(w,q)\\ \theta(w,q)\end{bmatrix} + \mathbf{u},$$

where $\mathbf{\Delta Y}$ is an $n\times 1$ vector of $\{\Delta Y_{it_i}\}_{i=1}^n$, $\iota(w,q)$ and $\mathbf{D}(w,q)$ are vectors of $1'$s and $0'$s such that elements in $\iota(w,q)$ and $\mathbf{D}(w,q)$ respectively take value 1 if and only if i corresponds to $(W_{it_i}=w, Q_{it_i}=q)$ and $(D_{it_i}=1, W_{it_i}=w, Q_{it_i}=q,)$, and $\mathbf{u}$ is a vector of $\{u_{it_i}\}_{i=1}^n$. By construction, $[\iota(w,q) : \mathbf{D}(w,q)]$ is orthogonal to $[\iota(w',q') : \mathbf{D}(w',q')]$ for all $(w,q)\neq(w',q')$, so that an orthogonal projection of $[\iota(w',q') : \mathbf{D}(w',q')]$

onto the space spanned by the columns of $[\iota(w,q) : \mathbf{D}(w,q)]$ is an $n \times 2$ matrix of $0'$s. Thus, applying the partition regression result (Frisch and Waugh, 1933), the OLS estimator from estimating (B) is the same as the OLS estimator obtained from estimating,

$$\Delta Y_{it_i} = \beta(w,q) + \theta(w,q) \times D_{it_i} + u_{it_i},$$

when only observations of i's that correspond to $(W_{it_i} = w, Q_{it_i} = q)$ are used. In this case, the OLS estimator for $\theta(w,q)$ is the difference between the averages of the treatment and control values of the dependent variable (e.g., see Imbens and Rubin, 2015) which proves our claim. ■

C Average Treatment Effects

Table C.I: Average Treatment Effects on Mental Health by Treatment Intensity (Full Table 2)

	Reliable Recovery (0-1)		Reliable Improvement (0-1)		Reliable Deterioration (0-1)	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Low Intensity						
Treatment	0.440***	0.430***	0.368***	0.360***	-0.078***	-0.078***
	(0.005)	(0.005)	(0.004)	(0.004)	(0.002)	(0.002)
Number of Individuals	491,942	491,942	491,942	491,942	491,942	491,942
Treatment Group	245,433	245,433	245,433	245,433	245,433	245,433
Control Group	246,509	246,509	246,509	246,509	246,509	246,509
R Squared	0.216	0.284	0.138	0.179	0.020	0.053
Panel B: High Intensity						
Treatment	0.439***	0.429***	0.404***	0.393***	-0.084***	-0.084***
	(0.008)	(0.008)	(0.007)	(0.006)	(0.003)	(0.002)
Number of Individuals	275,990	275,990	275,990	275,990	275990	275990
Treatment Group	136,379	136,379	136,379	136,379	136379	136379
Control Group	139,611	139,611	139,611	139,611	139611	139611
R Squared	0.234	0.298	0.164	0.198	0.021	0.069
Panel C: Step Up (Low to High Intensity)						
Treatment	0.449***	0.435***	0.404***	0.385***	-0.095***	-0.090***
	(0.004)	(0.005)	(0.004)	(0.004)	(0.002)	(0.002)
Number of Individuals	388,136	388,136	388,136	388,136	388136	388136
Treatment Group	191,868	191,868	191,868	191,868	191868	191868
Control Group	196,268	196,268	196,268	196,268	196268	196268
R Squared	0.244	0.296	0.164	0.200	0.024	0.078
Panel D: Step Down (High to Low Intensity)						
Treatment	0.452***	0.443***	0.395***	0.379***	-0.087***	-0.084***
	(0.009)	(0.008)	(0.010)	(0.007)	(0.004)	(0.004)
Number of Individuals	44,396	44,396	44,396	44,396	44396	44396
Treatment Group	21,752	21,752	21,752	21,752	21752	21752
Control Group	22,644	22,644	22,644	22,644	22644	22644
R Squared	0.235	0.307	0.158	0.208	0.022	0.077
Panel E: Intensity Not Recorded						
Treatment	0.427***	0.426***	0.367***	0.371***	-0.088***	-0.095***
	(0.012)	(0.013)	(0.009)	(0.008)	(0.004)	(0.004)

Number of Individuals	46,328	46,328	46,328	46,328	46328	46328
Treatment Group	23,142	23,142	23,142	23,142	23142	23142
Control Group	23,186	23,186	23,186	23,186	23186	23186
R Squared	0.217	0.292	0.135	0.184	0.021	0.079
Therapy Controls	No	Yes	No	Yes	No	Yes
Individual Controls	No	Yes	No	Yes	No	Yes
Service Controls	No	Yes	No	Yes	No	Yes
Local-Area Controls	No	Yes	No	Yes	No	Yes
Service Fixed Effects	No	Yes	No	Yes	No	Yes
Time Fixed Effects	No	Yes	No	Yes	No	Yes

Note: Linear probability models. Binary dependent variables. Robust standard errors clustered at service level in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table C.II: Average Treatment Effects on Mental Health by Treatment Intensity

	Δ PHQ-9 (0-27)		Δ GAD-7 (0-21)		Δ Mental Health Index (Z-Score)	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Low Intensity						
Treatment	-4.579***	-4.514***	-4.488***	-4.409***	-0.732***	-0.720***
	(0.059)	(0.054)	(0.054)	(0.050)	(0.009)	(0.008)
Number of Individuals	491,942	491,942	491,942	491,942	491,942	491,942
Treatment Group	245,433	245,433	245,433	245,433	245,433	245,433
Control Group	246,509	246,509	246,509	246,509	246,509	
R Squared	0.147	0.274	0.166	0.271	0.187	0.313
Panel B: High Intensity						
Treatment	-5.458***	-5.486***	-5.047***	-5.035***	-0.846***	-0.847***
	(0.110)	(0.084)	(0.084)	(0.077)	(0.015)	(0.013)
Number of Individuals	275,990	275,990	275,990	275,990	275,990	275,990
Treatment Group	136,379	136,379	136,379	136,379	136,379	136,379
Control Group	139,611	139,611	139,611	139,611	139,611	139,611
R Squared	0.186	0.291	0.196	0.283	0.223	0.329
Panel C: Step Up (Low to High Intensity)						
Treatment	-5.879***	-5.662***	-5.422***	-5.161***	-0.910***	-0.090***
	(0.063)	(0.060)	(0.051)	(0.049)	(0.009)	(0.002)
Number of Individuals	388,136	388,136	388,136	388,136	388,136	388,136
Treatment Group	191,868	191,868	191,868	191,868	191,868	191,868
Control Group	196,268	196,268	196,268	196,268	196,268	196,268
R Squared	0.199	0.309	0.210	0.304	0.237	0.078
Panel D: Step Down (High to Low Intensity)						
Treatment	-5.359***	-5.235***	-5.105***	-4.937***	-0.844***	-0.820***
	(0.180)	(0.147)	(0.150)	(0.120)	(0.026)	(0.021)
Number of Individuals	44,396	44,396	44,396	44,396	44,396	44,396
Treatment Group	21,752	21,752	21,752	21,752	21,752	21,752
Control Group	22,644	22,644	22,644	22,644	22,644	22,644
R Squared	0.175	0.311	0.193	0.305	0.215	0.351
Panel E: Intensity Not Recorded						
Treatment	-5.147***	-5.338***	-4.752***	-4.893***	-0.797***	-0.823***
	(0.114)	(0.128)	(0.108)	(0.123)	(0.017)	(0.020)
Number of Individuals	46,328	46,328	46,328	46,328	46,328	46,328
Treatment Group	23,142	23,142	23,142	23,142	23,142	23,142
Control Group	23,186	23,186	23,186	23,186	23,186	23,186

R Squared	0.160	0.282	0.168	0.274	0.191	0.317
Therapy Controls	No	Yes	No	Yes	No	Yes
Individual Controls	No	Yes	No	Yes	No	Yes
Service Controls	No	Yes	No	Yes	No	Yes
Local-Area Controls	No	Yes	No	Yes	No	Yes
Service Fixed Effects	No	Yes	No	Yes	No	Yes
Time Fixed Effects	No	Yes	No	Yes	No	Yes

Note: Robust standard errors clustered at service level in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table C.III: Average Treatment Effects: Robustness – Other Percentiles of Waiting Time

	Reliable Recovery (0-1)		Reliable Improvement (0-1)		Reliable Deterioration (0-1)	
	(1)	(2)	(3)	(4)	(5)	(6)
25th Percentile of Waiting Time						
Treatment	0.443***	0.458***	0.402***	0.419***	-0.079***	-0.076***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.002)	(0.001)
Number of Individuals	1,246,792	1,246,792	1,246,792	1,246,792	1,246,792	1,246,792
Treatment Group	294,571	294,571	294,571	294,571	294,571	294,571
Control Group	952,221	952,221	952,221	952,221	952,221	952,221
R Squared	0.228	0.280	0.119	0.148	0.011	0.062
75th Percentile of Waiting Time						
Treatment	0.438***	0.464***	0.373***	0.396***	-0.092***	-0.093***
	(0.004)	(0.004)	(0.003)	(0.003)	(0.002)	(0.001)
Number of Individuals	1,246,792	1,246,792	1,246,792	1,246,792	1,246,792	1,246,792
Treatment Group	926,894	926,894	926,894	926,894	926,894	926,894
Control Group	319,898	319,898	319,898	319,898	319,898	319,898
R Squared	0.145	0.222	0.116	0.155	0.023	0.058
90th Percentile of Waiting Time						
Treatment	0.437***	0.456***	0.365***	0.385***	-0.097***	-0.095***
	(0.004)	(0.005)	(0.003)	(0.004)	(0.002)	(0.002)
Number of Individuals	1,246,792	1,246,792	1,246,792	1,246,792	1,246,792	1,246,792
Treatment Group	1,121,181	1,121,181	1,121,181	1,121,181	1,121,181	1,121,181
Control Group	125,611	125,611	125,611	125,611	125,611	125,611
R Squared	0.069	0.153	0.058	0.101	0.015	0.044
Therapy Controls	No	Yes	No	Yes	No	Yes
Individual Controls	No	Yes	No	Yes	No	Yes
Service Controls	No	Yes	No	Yes	No	Yes
Local-Area Controls	No	Yes	No	Yes	No	Yes
Service Fixed Effects	No	Yes	No	Yes	No	Yes
Time Fixed Effects	No	Yes	No	Yes	No	Yes

Note: Linear probability models. Binary dependent variables. Robust standard errors clustered at service level in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table C.IV: Average Treatment Effects: Robustness – Other Models and Outcomes

	Reliable Recovery (0-1)			Other Outcomes		
	Logit Marginal Effect	Without Substance Abuse	Only Depression, Anxiety	Δ PHQ-9 (0-27)	Δ GAD-7 (0-21)	Δ Mental Health Index (Z-Score)
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.381***	0.431***	0.431***	-5.126***	-4.808***	-0.800***
	(0.003)	(0.004)	(0.004)	(0.052)	(0.044)	(0.008)
Therapy Controls	Yes	Yes	Yes	Yes	Yes	Yes
Individual Controls	Yes	Yes	Yes	Yes	Yes	Yes
Service Controls	Yes	Yes	Yes	Yes	Yes	Yes
Local-Area Controls	Yes	Yes	Yes	Yes	Yes	Yes
Service Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Number of Individuals	1,246,729	1,246,155	996,358	1,246,792	1,246,792	1,246,792
Treatment Group	618,521	618,239	491,358	618,574	618,574	618,574
Control Group	628,208	627,916	504,761	628,218	628,218	628,218
(Pseudo) R Squared	0.263	0.289	0.290	0.286	0.281	0.324

Note: Robust standard errors clustered at service level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table C.V: Average Treatment Effects on Work and Social Functioning

	Work and Social Adjustment Scale					
	Δ Overall (0-40)	Δ Work (0-8)	Δ Home Management (0-8)	Δ Social Leisure (0-8)	Δ Private Leisure (0-8)	Δ Close Relationships (0-8)
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	-5.709***	-1.091***	-0.998***	-1.390***	-1.084***	-1.145***
	(0.079)	(0.019)	(0.016)	(0.017)	(0.017)	(0.017)
Therapy Controls	Yes	Yes	Yes	Yes	Yes	Yes
Individual Controls	Yes	Yes	Yes	Yes	Yes	Yes
Service Controls	Yes	Yes	Yes	Yes	Yes	Yes
Local-Area Controls	Yes	Yes	Yes	Yes	Yes	Yes
Service Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Number of Individuals	750,351	750,351	750,351	750,351	750,351	750,351
Treatment Group	369,506	369,506	369,506	369,506	369,506	369,506
Control Group	380,845	380,845	380,845	380,845	380,845	380,845
R Squared	0.138	0.069	0.068	0.104	0.072	0.074

Note: Robust standard errors clustered at service level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table C.VI: Average Treatment Effects on Employment and Benefits

	Employed (*vs.* Unemployed)		Employed (*vs.* Long-Term Sick)		Receiving Statutory Sick Pay	
	Average	If Unemployed At Baseline	Average	If LT Sick At Baseline	Average	If St. Sick Pay at Baseline
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.001	0.029***	0.004***	0.023***	-0.005***	-0.032***
	(0.001)	(0.004)	(0.001)	(0.006)	(0.001)	(0.004)
Pre-Treatment Outcome	Yes	No	Yes	No	Yes	No
Therapy Controls	Yes	Yes	Yes	Yes	Yes	Yes
Individual Controls	Yes	Yes	Yes	Yes	Yes	Yes
Service Controls	Yes	Yes	Yes	Yes	Yes	Yes
Local-Area Controls	Yes	Yes	Yes	Yes	Yes	Yes
Service Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Number of Individuals	721,523	80,137	694,187	63,546	1,081,196	83,000
Treatment Group	359,089	39,993	340,429	27,872	531,560	44,331
Control Group	362,434	40,144	353,758	35,674	549,636	38,669
R Squared	0.549	0.106	0.767	0.079	0.106	0.101

Note: Linear probability models. Binary dependent variables. Robust standard errors clustered at service level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

D Robustness Checks: Attrition

Our primary analysis includes patients who attended at least three sessions, including an initial assessment session. During the initial assessment, the therapist and the patient decide whether the patient should continue with treatment in the programme. Patients unsuitable for IAPT treatment are referred to other services. Those within the program's scope can choose not to participate. In this section, our focus is on patients who were accepted into the program, agreed to participate, but subsequently dropped out before the second session, totaling 260,200 patients.

If attrition is selective, i.e. the probability of dropping out is correlated with the probability of recovery, it can bias our treatment effect estimates. Since we do not observe these patients after the first session, we lack information on whether their condition improved or deteriorated. We investigate potential impact of attrition on our programme effectiveness estimates by assuming various recovery rates for this group.

We impute the waiting time for patients who dropped out based on the average waiting time for the treatment intensity they were assigned to at the service they attended in the month of assessment. Subsequently, based on their waiting time, we allocate them to the treatment or control group using the same thresholds as in our main results.[41]

To bounds the estimates for three main outcomes (reliable recovery, reliable improvement, and reliable deterioration), we consider four scenarios:

- *Scenario 1:* All patients who dropped out of the treatment group deteriorated; hence, none recovered. All patients who dropped out of the control group improved and recovered, none deteriorated. This scenario provides an extreme lower bound for the treatment effect estimate because it elevates natural recovery rates estimated on the control group and suppresses recovery rates at the end of the program, estimated on the treatment group.

- *Scenario 2:* All patients who dropped out of the treatment and the control group improved and recovered, none deteriorated.

- *Scenario 3:* All patients who dropped out of the treatment and the control group deteriorated, and none improved or recovered.

- *Scenario 4:* All patients who dropped out of the treatment group improved and recovered, and none deteriorated. All patients who dropped out of the control group deteriorated; hence, none recovered. This scenario is the opposite of the first option and provides an extreme upper bound.

Table D.I reports the outcomes of models that include all controls for the four specified scenarios. Column 1 presents the main results for the reference. Across all scenarios, the programme significantly increases the

[41] Patients who drop out are typically located in services with longer waiting times; 74.56% of them were assigned to the control group. They are more likely to receive low-intensity treatment, 67.07% compared to 39.46% in the main sample. The symptoms of low-intensity patients who dropped out are slightly more severe than in the main sample, whereas symptoms are slightly less severe for other treatment intensities.

probability of recovery and improvement. Additionally, in all scenarios except the most extreme Scenario 1, the programme significantly reduces the probability of deterioration.

Table D.I: Average Treatment Effects on Mental Health for Different Recovery Scenarios of Drop-Out Patients

	Main result Table 1	Scenario 1	Scenario 2	Scenario 3	Scenario 4
	(1)	(2)	(3)	(4)	(5)
Reliable Recovery					
Treatment	0.431***	0.218***	0.296***	0.404***	0.483***
	(0.004)	(0.009)	(0.007)	(0.004)	(0.004)
R Squared	0.29	0.10	0.16	0.27	0.36
Reliable Improvement					
Treatment	0.377***	0.195***	0.273***	0.381***	0.460***
	(0.003)	(0.008)	(0.005)	(0.005)	(0.004)
R Squared	0.19	0.07	0.12	0.21	0.28
Reliable Deterioration					
Treatment	-0.084***	0.016***	-0.063***	-0.171***	-0.249***
	(0.001)	(0.005)	(0.001)	(0.007)	(0.007)
R Squared	0.06	0.06	0.05	0.16	0.21
Number of Individuals	1,246,792	1,507,012	1,507,012	1,507,012	1,507,012
Treatment Group	628,218	684,786	684,786	684,786	684,786
Control Group	618,574	822,226	822,226	822,226	822,226

Note: Linear probability model with all controls. Binary dependent variables. Robust standard errors clustered at service level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

E Heterogeneous Treatment Effects

Table E.I: Descriptive statistics of for the full sample and the nonparametric estimation sample

	Full sample		Nonparametric sample	
	Mean	Standard Deviation	Mean	Standard Deviation
Outcomes				
Reliable recovery	0.312	0.463	0.309	0.462
Reliable improvement	0.549	0.498	0.546	0.498
Reliable deterioration	0.093	0.291	0.091	0.287
Covariates				
Course intensity: Low intensity	0.395	0.489	0.445	0.497
High intensity	0.221	0.415	0.215	0.411
Step down	0.036	0.185	0.007	0.083
Step up	0.311	0.463	0.328	0.469
Undefined	0.037	0.189	0.005	0.073
Severity above median	0.497	0.500	0.490	0.500
Long-term health condition	0.202	0.402	0.131	0.337
Religion: Christian	0.191	0.393	0.163	0.369
Not religious	0.328	0.470	0.347	0.476
Other religion and missing	0.481	0.500	0.490	0.500
Ethnicity: White British	0.632	0.482	0.637	0.481
Other	0.081	0.273	0.017	0.128
Missing	0.287	0.452	0.347	0.476
Deprivation above median	0.551	0.497	0.551	0.497
Service size above median (number of staff)	0.500	0.500	0.506	0.500
Service funding per patient above median	0.499	0.500	0.514	0.500
Months: 2 or less	0.380	0.485	0.441	0.496
3	0.213	0.409	0.229	0.420
4	0.132	0.339	0.125	0.330
5	0.082	0.275	0.065	0.246
6	0.053	0.223	0.026	0.160
7 or above	0.140	0.347	0.115	0.319
Observations	1,246,792		947,547	

Table E.II: Heterogeneous treatment effect estimates. Full result for Table 5.3.

	Reliable recovery	Reliable improvement	Reliable deterioration
Treated	0.461***	0.371***	-0.099***
	(0.003)	(0.003)	(0.002)
Course intensity: Low intensity	0	0	0
	(.)	(.)	(.)
High intensity	-0.030***	-0.054***	0.021***
	(0.002)	(0.002)	(0.001)
Step down	-0.001	-0.014*	0.006
	(0.007)	(0.008)	(0.005)
Step up	-0.040***	-0.063***	0.036***
	(0.002)	(0.002)	(0.001)
Undefined	-0.002	0.023**	0.024***
	(0.008)	(0.009)	(0.006)
Severity above median	-0.105***	0.103***	-0.131***
	(0.001)	(0.001)	(0.001)
Deprivation above median, 1 if true	-0.023***	-0.044***	0.026***
	(0.001)	(0.001)	(0.001)
Long-term health condition	-0.013***	-0.039***	0.016***
	(0.002)	(0.002)	(0.001)
Service size above median (number of staff)	-0.001	0.003**	-0.002***
	(0.001)	(0.001)	(0.001)
Service funding per patient above median	-0.006***	-0.022***	0.010***
	(0.001)	(0.001)	(0.001)
Christian	0	0	0
	(.)	(.)	(.)
Not religious	-0.014***	-0.006***	-0.002*
	(0.002)	(0.002)	(0.001)
Other religion and missing	-0.012***	-0.009***	0.001
	(0.002)	(0.003)	(0.002)
White	0	0	0
	(.)	(.)	(.)
Other	-0.006	-0.026***	0.031***
	(0.005)	(0.006)	(0.004)
Missing	0.006***	0.007***	0.002*
	(0.002)	(0.002)	(0.001)
Months: 2 or less	0	0	0
	(.)	(.)	(.)
3	0.011***	0.025***	0.013***
	(0.002)	(0.002)	(0.001)
4	0.013***	0.032***	0.019***
	(0.002)	(0.002)	(0.001)
5	0.012***	0.037***	0.021***
	(0.002)	(0.003)	(0.002)
6	0.017***	0.043***	0.020***
	(0.004)	(0.004)	(0.003)
7 or above	0.013***	0.047***	0.028***
	(0.002)	(0.002)	(0.001)

Low intensity # Treated	0	0	0
	(.)	(.)	(.)
High intensity # Treated	0.002	0.039***	-0.016***
	(0.002)	(0.003)	(0.002)
Step down # Treated	0.003	0.017	0.001
	(0.010)	(0.012)	(0.007)
Step up # Treated	-0.018***	0.021***	-0.019***
	(0.002)	(0.003)	(0.002)
Undefined # Treated	-0.036***	-0.066***	-0.011
	(0.012)	(0.013)	(0.008)
Severity above median # Treated	-0.088***	-0.071***	0.096***
	(0.002)	(0.002)	(0.001)
Deprivation above median, 1 if true # Treated	-0.027***	0.004**	-0.014***
	(0.002)	(0.002)	(0.001)
Long-term health condition # Treated	-0.026***	0.003	-0.008***
	(0.003)	(0.003)	(0.002)
Service size above median (number of staff) # Treated	-0.004**	-0.006***	0.003**
	(0.002)	(0.002)	(0.001)
Service funding per patient above median # Treated	0.021***	0.026***	-0.010***
	(0.002)	(0.002)	(0.001)
Christian # Treated	0	0	0
	(.)	(.)	(.)
Not religious # Treated	-0.025***	-0.013***	0.007***
	(0.003)	(0.003)	(0.002)
Other religion and missing # Treated	-0.030***	-0.021***	0.006***
	(0.003)	(0.004)	(0.002)
White # Treated	0	0	0
	(.)	(.)	(.)
Other # Treated	-0.018**	0	-0.016***
	(0.007)	(0.008)	(0.005)
Missing # Treated	-0.055***	-0.030***	0.002
	(0.003)	(0.003)	(0.002)
2 or less # Treated	0	0	0
	(.)	(.)	(.)
3 # Treated	0.111***	0.069***	-0.025***
	(0.002)	(0.003)	(0.002)
4 # Treated	0.129***	0.076***	-0.033***
	(0.003)	(0.003)	(0.002)
5 # Treated	0.125***	0.065***	-0.030***
	(0.003)	(0.004)	(0.002)
6 # Treated	0.132***	0.064***	-0.033***
	(0.005)	(0.006)	(0.004)
7 or above # Treated	0.115***	0.050***	-0.032***
	(0.003)	(0.003)	(0.002)
Constant	0.188***	0.368***	0.149***
	(0.002)	(0.002)	(0.001)
R2	0.26	0.16	0.05
Observations	947,547	947,547	947,547

Table E.III: Average values of covariates by quartiles of estimated treatment effects. Reliable recovery.

	1 quartile	2 quartile	3 quartile	4 quartile
Individual characteristics				
Age, standardised	-0.141	0.037	0.063	0.041
Ex-services member of armed forces	0.012	0.012	0.012	0.018
Not an ex-services member or their dependant	0.484	0.520	0.482	0.778
Dependant of an ex-services member	0.002	0.002	0.002	0.003
No Response (armed forces)	0.502	0.466	0.504	0.201
Employed	0.338	0.536	0.588	0.815
Unemployed and Seeking Work	0.186	0.093	0.100	0.002
Students FT	0.072	0.051	0.064	0.028
Long-term sick or disabled	0.193	0.107	0.010	0.000
Homemaker	0.065	0.051	0.053	0.025
Not receiving benefits and not working or searching	0.037	0.022	0.024	0.011
Unpaid voluntary work	0.004	0.004	0.004	0.002
Retired	0.028	0.073	0.093	0.087
No Response (employment)	0.078	0.062	0.064	0.029
White background	0.527	0.579	0.530	0.893
Mixed background	0.017	0.015	0.018	0.015
Asian background	0.041	0.030	0.037	0.028
Black background	0.023	0.018	0.021	0.016
Other background (ethnicity)	0.013	0.010	0.012	0.008
No Response (ethnicity)	0.378	0.348	0.381	0.040
Male	0.222	0.237	0.218	0.313
Female	0.435	0.450	0.438	0.663
Indeterminate gender	0.000	0.000	0.001	0.000
No Response (gender)	0.343	0.313	0.344	0.024
Long term health condition	0.214	0.185	0.178	0.231
No long term health condition	0.354	0.413	0.389	0.653
No Response (health condition)	0.432	0.401	0.433	0.116
Religion: Christian	0.155	0.168	0.169	0.269
Not religious	0.286	0.306	0.268	0.454
Other religion	0.060	0.047	0.054	0.055
No Response (religion)	0.498	0.479	0.509	0.222
Heterosexual or Straight	0.481	0.517	0.481	0.776
Gay or Lesbian	0.016	0.015	0.015	0.021
Bisexual	0.014	0.013	0.013	0.016
Other sexual orientation or not listed	0.010	0.009	0.009	0.008
No Response (sexual orientation)	0.480	0.447	0.482	0.178
Relative deprivation of patient postcode (by LSOA), std.	-0.203	0.026	0.067	0.111
Treatment characteristics				
Course intensity: Low intensity	0.400	0.489	0.362	0.327
Course intensity: High intensity	0.274	0.219	0.202	0.190
Course intensity: Step down	0.034	0.033	0.038	0.038
Course intensity: Step up	0.253	0.226	0.359	0.407
Course intensity: Undefined	0.038	0.033	0.039	0.038
Initial diagnosis: Anxiety and stress related disorders	0.010	0.007	0.006	0.004
Initial diagnosis: Depression	0.163	0.232	0.229	0.258
Initial diagnosis: Other problems	0.049	0.059	0.067	0.071

Initial diagnosis: Unspecified or Invalid Data	0.778	0.702	0.697	0.667
Medication usage: Prescribed but not taking	0.048	0.043	0.046	0.045
Medication usage: Prescribed and taking	0.557	0.454	0.446	0.449
Medication usage: Not Prescribed	0.322	0.432	0.446	0.460
No Response (medication usage)	0.073	0.071	0.062	0.045
Symptoms severity at start	0.998	0.209	0.265	0.263
Appointment month	-0.010	-0.014	-0.001	0.024
Referral type: Primary Health Care	0.240	0.219	0.223	0.185
Referral type: Self Referral	0.675	0.712	0.714	0.759
Referral type: Other	0.086	0.068	0.063	0.056
Treatment mode: Face to face communication	0.316	0.294	0.264	0.243
Treatment mode: Telephone	0.646	0.667	0.699	0.726
Treatment mode: Other	0.038	0.039	0.037	0.030
Appointment weekday	2.914	2.921	2.914	2.921
Service characteristics				
CCG Allocations per registered patient, standardised	0.026	-0.024	-0.037	0.035
CCG Estimated registered patients, standardised	-0.003	0.010	0.051	-0.058
CCG Number of Staff, standardised	0.007	0.002	0.021	-0.030
CCG Number of Staff, missing	0.055	0.049	0.048	0.061
Local area characteristics				
IMD: Crime - Average rank, standardised	0.052	-0.021	-0.008	-0.022
IMD: Education, Skills and Training - Average rank, std.	0.037	-0.041	-0.085	0.089
IMD: Employment - Average rank, standardised	0.046	-0.051	-0.075	0.080
IMD: Living Environment - Average rank, standardised	0.006	-0.002	0.039	-0.044
IMD: Health Deprivation and Disability - Average rank, std.	0.039	-0.042	-0.082	0.086
IMD: Barriers to Housing and Services - Average rank, std.	0.015	0.015	0.097	-0.128
IMD: Income - Average rank, standardised	0.053	-0.044	-0.039	0.031
IMD - Average rank, standardised	-0.049	0.042	0.043	-0.036
CCG Median Wage, standardised	-0.012	0.025	0.090	-0.103
CCG Unemployment Rate	4.429	4.321	4.325	4.392
Waiting times				
Months wait: 2 or less	0.736	0.673	0.112	0.000
Months wait: 3	0.080	0.110	0.316	0.345
Months wait: 4	0.053	0.066	0.190	0.219
Months wait: 5	0.036	0.044	0.117	0.132
Months wait: 6	0.024	0.029	0.074	0.083
Months wait: 7	0.017	0.020	0.049	0.057
Months wait: 8 or above	0.054	0.057	0.141	0.164

Table E.IV: Average values of covariates by quartiles of estimated treatment effects. Reliable improvement.

	1 quartile	2 quartile	3 quartile	4 quartile
Individual characteristics				
Age, standardised	-0.053	0.010	0.050	-0.007
Ex-services member of armed forces	0.011	0.012	0.014	0.016
Not an ex-services member or their dependant	0.437	0.522	0.555	0.749
Dependant of an ex-services member	0.002	0.002	0.003	0.003
No Response (armed forces)	0.550	0.464	0.428	0.232
Employed	0.547	0.554	0.576	0.600
Unemployed and Seeking Work	0.116	0.106	0.085	0.072
Students FT	0.054	0.053	0.051	0.058
Long-term sick or disabled	0.091	0.084	0.076	0.059
Homemaker	0.051	0.049	0.046	0.047
Not receiving benefits and not working or searching	0.026	0.024	0.023	0.020
Unpaid voluntary work	0.003	0.004	0.004	0.004
Retired	0.054	0.067	0.080	0.080
No Response (employment)	0.058	0.058	0.059	0.058
White background	0.484	0.566	0.623	0.856
Mixed background	0.014	0.019	0.015	0.017
Asian background	0.030	0.043	0.030	0.033
Black background	0.018	0.026	0.016	0.019
Other background (ethnicity)	0.009	0.014	0.009	0.010
No Response (ethnicity)	0.445	0.332	0.306	0.064
Male	0.197	0.237	0.242	0.313
Female	0.391	0.468	0.485	0.641
Indeterminate gender	0.000	0.000	0.001	0.001
No Response (gender)	0.412	0.294	0.272	0.045
Long term health condition	0.166	0.201	0.198	0.243
No long term health condition	0.346	0.417	0.438	0.609
No Response (health condition)	0.488	0.382	0.364	0.148
Religion: Christian	0.142	0.171	0.193	0.256
Not religious	0.260	0.299	0.319	0.436
Other religion	0.045	0.059	0.051	0.062
No Response (religion)	0.553	0.471	0.437	0.247
Heterosexual or Straight	0.436	0.524	0.550	0.744
Gay or Lesbian	0.014	0.017	0.015	0.020
Bisexual	0.012	0.014	0.013	0.017
Other sexual orientation or not listed	0.007	0.010	0.009	0.010
No Response (sexual orientation)	0.531	0.434	0.413	0.208
Relative deprivation of patient postcode (by LSOA), std.	-0.015	0.019	-0.022	0.018
Treatment characteristics				
Course intensity: Low intensity	0.491	0.416	0.350	0.321
Course intensity: High intensity	0.248	0.213	0.220	0.204
Course intensity: Step down	0.030	0.033	0.038	0.042
Course intensity: Step up	0.198	0.302	0.353	0.393
Course intensity: Undefined	0.033	0.036	0.039	0.041
Initial diagnosis: Anxiety and stress related disorders	0.008	0.008	0.006	0.006
Initial diagnosis: Depression	0.203	0.213	0.228	0.237
Initial diagnosis: Other problems	0.049	0.059	0.064	0.074

Initial diagnosis: Unspecified or Invalid Data	0.740	0.720	0.701	0.683
Medication usage: Prescribed but not taking	0.048	0.048	0.045	0.042
Medication usage: Prescribed and taking	0.523	0.487	0.465	0.432
Medication usage: Not Prescribed	0.364	0.399	0.429	0.467
No Response (medication usage)	0.065	0.066	0.062	0.058
Symptoms severity at start	0.887	0.566	0.328	-0.046
Appointment month	-0.002	-0.007	-0.003	0.012
Referral type: Primary Health Care	0.226	0.229	0.214	0.198
Referral type: Self Referral	0.707	0.704	0.716	0.734
Referral type: Other	0.067	0.067	0.070	0.068
Treatment mode: Face to face communication	0.291	0.276	0.281	0.269
Treatment mode: Telephone	0.668	0.685	0.687	0.698
Treatment mode: Other	0.041	0.039	0.032	0.033
Appointment weekday	2.922	2.913	2.919	2.915
Service characteristics				
CCG Allocations per registered patient, standardised	-0.062	-0.090	0.066	0.085
CCG Estimated registered patients, standardised	0.069	0.116	-0.105	-0.080
CCG Number of Staff, standardised	0.036	0.042	-0.055	-0.024
CCG Number of Staff, missing	0.047	0.038	0.063	0.066
Local area characteristics				
IMD: Crime - Average rank, standardised	-0.073	-0.045	0.057	0.061
IMD: Education, Skills and Training - Average rank, std.	-0.117	-0.154	0.119	0.153
IMD: Employment - Average rank, standardised	-0.153	-0.173	0.142	0.183
IMD: Living Environment - Average rank, standardised	-0.042	0.082	-0.024	-0.016
IMD: Health Deprivation and Disability - Average rank, std.	-0.153	-0.173	0.141	0.185
IMD: Barriers to Housing and Services - Average rank, std.	0.086	0.188	-0.126	-0.148
IMD: Income - Average rank, standardised	-0.128	-0.109	0.104	0.134
IMD - Average rank, standardised	0.127	0.108	-0.101	-0.134
CCG Median Wage, standardised	0.095	0.139	-0.090	-0.144
CCG Unemployment Rate	4.234	4.229	4.491	4.513
Waiting times				
Months wait: 2 or less	0.873	0.434	0.215	0.000
Months wait: 3	0.046	0.183	0.270	0.353
Months wait: 4	0.022	0.112	0.170	0.223
Months wait: 5	0.018	0.077	0.104	0.130
Months wait: 6	0.012	0.051	0.066	0.082
Months wait: 7	0.008	0.034	0.044	0.056
Months wait: 8 or above	0.021	0.109	0.130	0.157

Table E.V: Average values of covariates by quartiles of estimated treatment effects. Reliable deterioration.

	1 quartile	2 quartile	3 quartile	4 quartile
Individual characteristics				
Age, standardised	0.029	-0.005	-0.043	0.019
Ex-services member of armed forces	0.013	0.012	0.013	0.015
Not an ex-services member or their dependant	0.592	0.540	0.569	0.562
Dependant of an ex-services member	0.002	0.002	0.003	0.003
No Response (armed forces)	0.393	0.445	0.415	0.421
Employed	0.600	0.613	0.571	0.493
Unemployed and Seeking Work	0.073	0.075	0.098	0.134
Students FT	0.058	0.059	0.057	0.042
Long-term sick or disabled	0.049	0.047	0.079	0.134
Homemaker	0.048	0.045	0.049	0.052
Not receiving benefits and not working or searching	0.019	0.018	0.024	0.033
Unpaid voluntary work	0.004	0.004	0.003	0.003
Retired	0.088	0.083	0.062	0.049
No Response (employment)	0.061	0.056	0.056	0.060
White background	0.666	0.608	0.632	0.624
Mixed background	0.017	0.015	0.016	0.017
Asian background	0.033	0.027	0.034	0.042
Black background	0.020	0.017	0.020	0.022
Other background (ethnicity)	0.010	0.009	0.011	0.013
No Response (ethnicity)	0.255	0.323	0.286	0.283
Male	0.264	0.242	0.240	0.244
Female	0.512	0.466	0.503	0.504
Indeterminate gender	0.001	0.000	0.001	0.000
No Response (gender)	0.224	0.291	0.256	0.252
Long term health condition	0.195	0.174	0.202	0.238
No long term health condition	0.484	0.450	0.455	0.421
No Response (health condition)	0.321	0.376	0.343	0.341
Religion: Christian	0.204	0.181	0.187	0.189
Not religious	0.335	0.316	0.335	0.328
Other religion	0.053	0.045	0.054	0.064
No Response (religion)	0.408	0.457	0.424	0.418
Heterosexual or Straight	0.589	0.540	0.562	0.564
Gay or Lesbian	0.016	0.015	0.017	0.018
Bisexual	0.013	0.013	0.015	0.014
Other sexual orientation or not listed	0.009	0.008	0.009	0.010
No Response (sexual orientation)	0.372	0.424	0.397	0.395
Relative deprivation of patient postcode (by LSOA), std.	0.021	0.142	-0.012	-0.151
Treatment characteristics				
Course intensity: Low intensity	0.340	0.522	0.389	0.328
Course intensity: High intensity	0.208	0.198	0.219	0.260
Course intensity: Step down	0.042	0.031	0.034	0.036
Course intensity: Step up	0.368	0.219	0.322	0.337
Course intensity: Undefined	0.042	0.031	0.036	0.040
Initial diagnosis: Anxiety and stress related disorders	0.006	0.006	0.007	0.009
Initial diagnosis: Depression	0.226	0.259	0.224	0.174
Initial diagnosis: Other problems	0.077	0.059	0.057	0.053

Initial diagnosis: Unspecified or Invalid Data	0.691	0.676	0.712	0.764
Medication usage: Prescribed but not taking	0.042	0.043	0.048	0.050
Medication usage: Prescribed and taking	0.403	0.415	0.498	0.590
Medication usage: Not Prescribed	0.492	0.474	0.394	0.299
No Response (medication usage)	0.063	0.068	0.060	0.061
Symptoms severity at start	-0.278	0.023	0.698	1.292
Appointment month	0.004	-0.002	0.001	-0.003
Referral type: Primary Health Care	0.214	0.208	0.213	0.233
Referral type: Self Referral	0.718	0.733	0.720	0.690
Referral type: Other	0.068	0.060	0.067	0.078
Treatment mode: Face to face communication	0.277	0.273	0.269	0.299
Treatment mode: Telephone	0.688	0.685	0.697	0.668
Treatment mode: Other	0.035	0.042	0.035	0.033
Appointment weekday	2.919	2.922	2.914	2.915
Service characteristics				
CCG Allocations per registered patient, standardised	0.023	-0.062	0.007	0.032
CCG Estimated registered patients, standardised	-0.002	0.006	-0.005	0.001
CCG Number of Staff, standardised	0.001	-0.009	-0.003	0.010
CCG Number of Staff, missing	0.050	0.049	0.054	0.061
Local area characteristics				
IMD: Crime - Average rank, standardised	0.089	-0.075	-0.011	-0.002
IMD: Education, Skills and Training - Average rank, std.	0.049	-0.118	0.007	0.063
IMD: Employment - Average rank, standardised	0.086	-0.134	0.000	0.048
IMD: Living Environment - Average rank, standardised	0.084	-0.027	-0.017	-0.040
IMD: Health Deprivation and Disability - Average rank, std.	0.080	-0.116	0.002	0.034
IMD: Barriers to Housing and Services - Average rank, std.	0.023	0.020	-0.024	-0.018
IMD: Income - Average rank, standardised	0.102	-0.124	-0.008	0.030
IMD - Average rank, standardised	-0.102	0.121	0.007	-0.026
CCG Median Wage, standardised	-0.011	0.080	-0.012	-0.057
CCG Unemployment Rate	4.481	4.239	4.354	4.394
Waiting times				
Months wait: 2 or less	0.073	0.613	0.396	0.439
Months wait: 3	0.336	0.138	0.199	0.178
Months wait: 4	0.204	0.083	0.129	0.112
Months wait: 5	0.122	0.051	0.080	0.075
Months wait: 6	0.076	0.032	0.052	0.050
Months wait: 7	0.051	0.021	0.035	0.035
Months wait: 8 or above	0.138	0.060	0.108	0.110