Talking Therapy: Impacts of a Nationwide Mental Health Service*

Ekaterina Oparina† LSE ⓡ

Christian Krekel‡ LSE ⓡ

Sorawoot Srisuma§ NUS ⓡ

March 2026

Abstract

Mental health problems impose high costs, yet healthcare systems often overlook them. We provide the first causal evidence on the effectiveness of a nationwide-scaled mental health service in England for treating depression and anxiety using non-experimental data and methods. Exploiting oversubscription and resulting exogenous variation in waiting times across areas and over time, and based on a novel dataset of over one million patients, we find that treatment significantly improves mental health and reduces impairment in work and social life. We also provide suggestive evidence of employment gains. Impacts vary across patients and services. Nevertheless, the programme is highly cost-effective.

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

JEL Codes: I18, D04, D61

*The symbol ⓡ indicates that the author order was determined randomly. We are grateful to Richard Layard and David Clark for valuable discussions throughout this project. Niall Maher and Isaac Parkes provided excellent research assistance. We thank Peter Coyte, Erzo Luttmer, Steve Machin, Guy Michaels, Fiona Scott Morton, Martin Knapp, Henry Overman, Steve Pilling, Imran Rasul, and participants at the Annual Meetings of the Allied Social Science Association, the Royal Economic Society, and the CEP Annual Conference, as well as seminar participants at the LSE, Mental Health and Economic Status Conference (University of Warwick), Office for National Statistics (ONS), Highland Health Economics Symposium, IFS-CAGE Workshop on the Economics of Mental Health, and Warwick Business School for comments and suggestions. This work was supported by the ESRC [Grant Number: ES/W002094/1].

†Centre for Economic Performance (CEP), London School of Economics (LSE) (e.oparina@lse.ac.uk)

‡CEP, LSE; Department of Psychological and Behavioural Science, LSE (c.krekel@lse.ac.uk), Corresponding Author

§Department of Economics, National University of Singapore (NUS) (s.srisuma@nus.edu.sg)

1 Introduction

Nearly one billion people globally live with a mental health disorder (WHO, 2022). The economic burden of mental ill health is estimated to reach $5 trillion, representing between 4% and 8% of GDP across different regions (Arias et al., 2022). Poor mental health is linked to worse labour market outcomes (Banerjee et al., 2017; Chatterji et al., 2011; Frijters et al., 2014) and educational attainment (J. M. Fletcher, 2010), with spillovers to families amplifying societal costs (J. Fletcher, 2009). Allocating resources to cost-effective mental health policies is key, as improvements in mental health enhance human capital leading to long-term economic benefits (Layard, 2016).

Despite the burden imposed by mental ill health, evidence on the effectiveness of population-wide mental health services is scarce. While randomised controlled trials (RCTs) provide strong evidence for the effectiveness of psychological therapies in highly controlled settings (Lambert, 2013; Nathan & Gorman, 2015; A. Roth & Fonagy, 2005), there are no guarantees that scaling these interventions to national levels will produce similar outcomes (List, 2022). This is due to larger scale, access to more diverse population groups, and the fact that patients choose to get treatment rather than being allocated to it (see Cronin et al., 2024, for a discussion on the importance of the latter in the context of public policy more generally).

This paper is the first to estimate the causal effects of a nationwide-scaled mental health service that provides psychological therapies to the general population.[1] We study its overall effect as well as heterogeneous effects across patients, services, and areas. Our results serve as a guide and benchmark for implementing similar policies worldwide, as countries increasingly recognise the growing economic burden of mental ill health and set up similar services.

Our work advances a rapidly growing economics literature on mental health treatment. A recent body of work has begun to provide causal evidence on the effects of mental health interventions in high-income countries, studying pharmacological treatments (Bhalotra et al., 2025; Biasi et al., 2026), guideline adherence

[1] An important contribution closest in spirit to ours is Serena (2025), who examines the impacts of expanding partial health insurance coverage of psychological therapy in Denmark on health service use, labour market outcomes, and suicide attempts. Serena's study covers the entire Danish population but focuses on patients aged 18 to 38 accessing private-practice psychologists who treat mild to moderate depression and anxiety, an institutional setup similar to ours. We examine the direct mental health impacts on all patients accessing clinicians trained within a public programme and adhering to its guidelines.

in prescribing (Cuddy & Currie, 2026), and the consequences of treatment delays (Costantini, 2025). A parallel literature, largely based on RCTs, documents positive impacts of CBT on a range of health and human capital outcomes, including perinatal depression and subsequent female empowerment and investments into children's cognitive and socio-emotional skills (Baranov et al., 2020; Sevim et al., 2023a, 2023b); mental health of individuals living in poor households (Barker et al., 2022); anti-social and criminal behaviour amongst economically disadvantaged youth (Blattman et al., 2017; Heller et al., 2017) or violence amongst prisoners (Batistich et al., 2024); self-image (Ghosal et al., 2022); and overall psychological and economic wellbeing (Bossuroy et al., 2022; Haushofer et al., 2022). Angelucci and Bennett (2024) look at antidepressants and livelihoods support, detecting impacts on mental health when combined. Most of these studies find medium to strong impacts that are often lasting.[2] Most of this evidence comes from developing countries. Notable exceptions include Heller et al. (2017), who study the impact of CBT on criminal behaviour in Chicago, and Batistich et al. (2024), who evaluate CBT for reducing violence amongst prisoners in Texas. These studies typically rely on small samples, limiting the scope for exploring heterogeneity or drawing general conclusions about programme effectiveness at scale.

Our paper makes three contributions to this literature. First, we provide the first causal estimates of a nationwide-scaled psychological therapy programme, complementing causal evidence from pharmacological treatments (Bhalotra et al., 2025; Biasi et al., 2026) and from smaller-scale therapy RCTs. Our quasi-experimental approach uses data on over one million patients – the universe of those treated – which can be useful for guiding counterfactual questions on scaling up smaller pilots to the policy level (cf. List, 2022). Second, our unique session-by-session patient-level outcome data allow us to separate treatment effects from natural recovery, which we show is the prevailing factor that makes estimates from non-causal before-after comparisons generally larger than actual treatment effects. Third, we provide novel evidence on heterogeneous treatment effects using both a non-parametric matching approach and generalised random forests, identifying employment status and self-referral as important sources of heterogeneity that have been overlooked in earlier studies.

The features of the programme we study make our findings relevant beyond

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

England. We study the *Improving Access to Psychological Therapies (IAPT)* programme, a nationwide mental health service in England that provides evidence-based psychological therapies for common mental health disorders, in particular depression and anxiety.[3] The programme is the largest in the world: to date, it has deployed over 10,500 new therapists and treated over seven million patients (more than 13% of the English population), primarily via CBT, also referred to as *talking therapies*. IAPT therapists are trained via a standardised national curriculum to provide psychological therapies recommended by the *National Institute for Health and Care Excellence (NICE)* in the UK and, hence, supported by an extensive body of causal evidence on their effectiveness. All therapists who work in the programme adhere to the same national treatment guidelines. IAPT services are free of charge to patients. These design features – standardised training, adherence to national clinical guidelines, free access, and a stepped-care model – are shared by IAPT-inspired programmes now being implemented in Australia, Canada (Ontario), Lithuania, Norway, Spain, and Sweden, making our findings directly informative for these contexts. The patients in our sample attended, on average, about seven sessions over a period of 14 weeks.

We use data on over one million patients from a novel dataset comprised of all individuals who started their treatment between April 2016 and December 2018.[4] Our main outcomes are binary indicators for *reliable recovery*, *reliable improvement*, and *reliable deterioration*, which are based on validated measures used by clinicians around the world to diagnose depression (PHQ-9 scores) and anxiety (GAD-7 scores), and which patients are required to complete before each session.[5] We also look at PHQ-9 and GAD-7 scores as well as a mental health index separately. Beyond mental health outcomes, we study work and social functioning in various domains as well as self-reported employment and statutory sick pay receipt. Our dataset is of an exceptionally high quality with outcomes recorded for 98% of patients. It is worldwide unique, in that it records patient-level outcomes on a session-by-session basis before each session start, offering precisely the variation we need to estimate

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

[4]The rollout started in 2008, and by 2016 the programme's operations and current outcome monitoring system were fully established.

[5]Reliable improvement is one if PHQ-9 and/or GAD-7 scores decreased by a reliable amount and neither score increased; reliable deterioration is one if PHQ-9 and/or GAD-7 scores increased by a reliable amount and neither decreased; reliable recovery is one if a patient reliably improved and both scores are below the clinical cut-off at the end of treatment. Section 3 provides more details.

causal effects, including specific features such as session value added.

We estimate causal effects using a quasi-experimental approach. We rely on oversubscription of patients to the programme for identification, which creates exogenous variation in waiting times, as more patients are referred to therapies than can be instantly treated.[6] This variation in waiting times differs across services (and, thereby, local areas) and over time depending on supply-side constraints, in particular shortages of trained therapists, and certain demand-side characteristics, in particular local clusters of mental ill health and regional differences in the balance between high-intensity and low-intensity treatment provision. We routinely control for treatment intensity as well as initial diagnosis and severity of symptoms at the start of treatment throughout our regressions; our results are robust to omitting these controls. We compare the change in mental health of patients who completed treatment (treatment group) to that of patients who were waiting for the start of treatment during the same period of time (control group), in a difference-in-differences design estimated as a first differences model, controlling for psychological-therapy, individual, service and associated local-area characteristics, as well as service and time fixed effects. Treatment in the programme is allocated strictly on a first-come, first-serve basis, a legal requirement based on fairness principles. In Section 4.1.1, we provide empirical evidence that waiting times are not systematically related to the severity of depression or anxiety symptoms at the start or end of treatment, the number of sessions, or treatment duration, neither across nor in any of the different intensity provisions of the programme.

Our empirical analysis leads to six key findings. First, we find that the programme causally improves mental health: treated patients are about 43 percentage points more likely to reliably recover than waitlisted patients, with reductions in depression symptoms (PHQ-9) of 5.1 points and anxiety symptoms (GAD-7) of 4.8 points. In our data, 54% of treated patients reliably recovered by the end of treatment, compared to 9% who naturally recovered in the waitlist control group. These recovery rates are similar to findings from RCTs of IAPT-style programmes in Norway (59%, Knapstad et al., 2020) and Spain (50%, Cano-Vindel et al., 2022), though

[6] The use of waitlists to identify treatment effects in economics is not new. An early contribution is found in Berger and Black (1992). The idea has also been implemented in experimental settings (cf. Finkelstein et al., 2019; Jacob & Ludwig, 2012; Jacob et al., 2015). More recent works, like ours, exploit naturally occurring waitlists due to oversubscription or excess demand (Beam & Quimbo, 2023; Dague et al., 2017; Dinerstein et al., 2022; Hoe, 2023; Robles et al., 2021). Costantini (2025) similarly exploits clinic congestion and resulting variation in waiting times in the context of mental health services in the Department of Veteran Affairs in the US.

unlike our study, these trials used treatment-as-usual control groups rather than waitlists, leading to higher control-group recovery rates. Our session-by-session outcome data show a steady session value added from the first to the last clinical session. Second, we detect positive short-term ripple effects on work and social life: amongst those initially unemployed or on long-term sick leave, treated patients are about three percentage points more likely to report being employed and three percentage points less likely to receive statutory sick pay at the end of treatment. Treated patients also report 65% of a standard deviation less functional impairment across all measured domains of life. Third, treated patients are significantly less likely to experience mental health deterioration, providing novel evidence that addresses concerns that psychological interventions may inadvertently cause harm (Harvey et al., 2023). Fourth, we find substantial heterogeneity in the programme's effectiveness. Patients typically at risk of poorer mental health outcomes – e.g. those living with a disability or residing in deprived areas – generally benefit less, while higher service funding is positively associated with outcomes. Using causal methods, we find that patients with long-term health conditions are approximately three percentage points less likely to reliably recover, which is significantly lower than the 14 percentage points difference estimated by Moller et al. (2019) using correlational methods, suggesting that a large part of the difference is due to natural recovery. Fifth, our machine-learning analysis identifies employment status as an important source of heterogeneity: unemployed patients are 13.3 percentage points less likely to recover as a result of treatment, representing 30% of the average treatment effect. Sixth, we find that self-referral – accessing treatment without a GP as a gatekeeper – improves access to care: self-referred patients sought care, on average, 364 days after the onset of symptoms versus 461 days for those referred via other pathways, and are 3.8 percentage points more likely to recover as a result of treatment.

These results are robust to different definitions of treatment and control group, to controlling for session spacing, to repeat enrolment, to different model specifications and estimation samples, and to alternative outcome measures. A bounding analysis shows that the programme remains effective even under extreme assumptions about selective dropout. To address potential selection on outcomes, we estimate treatment effects using data from the penultimate and antepenultimate clinical sessions and additionally control for the total number of attended sessions. Our results are remarkably robust to all of these checks. A conservative cost-benefit cal-

culation suggests that the benefits of the programme are at least five times larger than its costs.

2 The IAPT Programme

2.1 Institutional Context

In 2008, the UK Government launched the IAPT programme to make evidence-based psychological therapies more widely available within the National Health Service (NHS), focusing on the most common mental health problems: depression and anxiety disorders.[7] 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 at the time. CCGs were independent, geographically distinct bodies accountable to the Secretary of State for Health and Social Care through NHS England, each reflecting local needs and responsible for commissioning public healthcare for, on average, a quarter of a million of people NHS (2021b).[8] Today, IAPT is the largest programme of its kind in the world. It is seen as a pioneering model for treating mental ill health at the general population level, and is being replicated in other countries, e.g. Australia, Canada (Ontario), Lithuania, Norway, Spain, or Sweden.[9] By now, IAPT has treated over seven million patients (more than 13% of the English population) and the NHS has committed to further expand access (NHS, 2019).

The programme provides psychological therapies recommended by the *National Institute for Health and Care Excellence (NICE)* in the UK, an independent body

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

[8] CCGs emerged from *Primary Care Trusts (PCTs)* in 2013. In 2022, they were replaced with *Integrated Care Systems (ICS)*.

[9] The Norwegian adaptation is named *Prompt Mental Health Care* (*RPH* in Norwegian, see Knapstad et al. (2020) for a clinical trial). In Spain, Psicofundación developed the PsicAP clinical trial, following the IAPT approach (see Cano-Vindel et al. (2022)). Australia's *New Access* programme for depression and anxiety is strongly influenced by IAPT (see Baigent et al. (2023)).

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 CBT, and advocates a stepped-care model with both low and high-intensity treatments.[10] To access the programme, patients can either be referred by their GPs or they can refer themselves (so-called *self-referral*). The latter feature was a new option at the time the programme was launched, whose goal was to make psychological therapies more accessible amongst under-served population groups. There is universal public healthcare in England. As part of this, accessing the IAPT programme is free of charge to patients, without co-payment.

In their first session, patients undergo an initial assessment in which they are screened for the type of problem and the severity of symptoms. If patients are above the clinical caseness threshold for depression and/or anxiety, they are admitted and jointly agree with a trained therapist on a course of treatment; if they are below this threshold or their problem is considered more appropriate for a different service, they are signposted elsewhere.[11] Note that the IAPT programme was launched precisely because there was a lack of treatment options for mild to moderate cases of common mental disorders. Everybody who is admitted eventually gets treated. After the initial assessment, admitted patients are waitlisted and, after a while, start treatment in their second session, which constitutes the first clinical session. In most cases, the therapist who delivers the treatment is different from the therapist who performs the initial assessment. Those with mild to moderate symptoms start with low-intensity treatment (e.g. guided self-help or computerised CBT) and, if not responding, are upgraded to a higher intensity (usually weekly face-to-face one-to-one sessions); those with moderate to more severe symptoms, as well as with special forms of anxiety disorders such as post-traumatic stress disorder, start immediately with high-intensity treatment. If patients switch intensity, there will always be a switch in therapist too, as different therapists deliver low- and high-intensity treatment (see below for a detailed description of differences in education and training between low- and high-intensity therapists). While low-intensity treatment is often conducted online or over the phone as treatment mode,

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

[11] For example, the IAPT programme does not treat particularly severe cases or cases with complex co-morbidities.

high-intensity treatment is conducted in-person, delivered locally where patients and therapists live, typically at local GP practices or community centres rented by CCGs specifically for this purpose. Note that therapists only treat patients in their geographical area. About 60% of patients entering the programme (over 560,000 patients per year) receive at least one clinical session. Of these, the vast majority receive 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 (a change from low to high-intensity CBT), and a small number (8%) other forms of treatment (NHS, 2021a).[12] Therapy sessions are generally about 30 minutes for low- and 60 minutes for high-intensity treatment. There are no restrictions on the number of patients a therapist can see; in practice, the number of patients is determined by the contracted number of working hours a therapist has.[13]

CBT itself 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).[14] Therapists may prescribe medication additional to psychological therapy (which we are routinely controlling for in our regressions).

Specifically for the IAPT programme, the UK Department of Health and Social Care implemented a standardised training with dedicated national curricula for therapists covering a wide range of evidence-based CBT treatments.[15] New

[12]This small number of 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.

[13]Therapists have standard job contracts: if therapists are working full-time, they are expected to have about 20 hours per week of direct contact with patients. Because of the difference in the length of a therapy session between low- and high-intensity treatment, low-intensity therapists see about twice as many patients as high-intensity ones, though there is little variability in the number of patients *within* each category of therapist.

[14]Take a panic attack, for 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).

[15]These national curricula can be found at: https:\hee.nhs.uk. A competency framework, which

therapists working in the programme are required to learn at least two treatments for depression and one for each anxiety disorder. 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 to obtain an accredited postgraduate diploma (more days for trainees in high-intensity treatments, who are required to have prior experience in mental health services and are also paid more) and spend the rest of their time in on-the-job training. By 2019, about 10,500 newly trained therapists were deployed.[16]

Therapists are recruited and employed locally by each service (CCG), which determines staffing levels based on its allocated budget and local demand for psychological therapies. There is no centralised assignment of therapists to areas: each service recruits independently, and therapists treat only patients within their geographical area. Importantly, staff turnover and the pipeline of trainees completing the national curriculum are the main sources of variation in local therapist supply. The programme operates under national performance targets set by NHS England, most notably a target recovery rate of 50%. This target applies uniformly to all services and is monitored through the programme's mandatory outcome reporting system. It does not create incentives to prioritise particular patients over others: services are assessed on aggregate recovery rates across all treated patients, not on the outcomes of individual cases. There are no financial incentives tied to individual patient outcomes or to the allocation of patients between low- and high-intensity treatment.

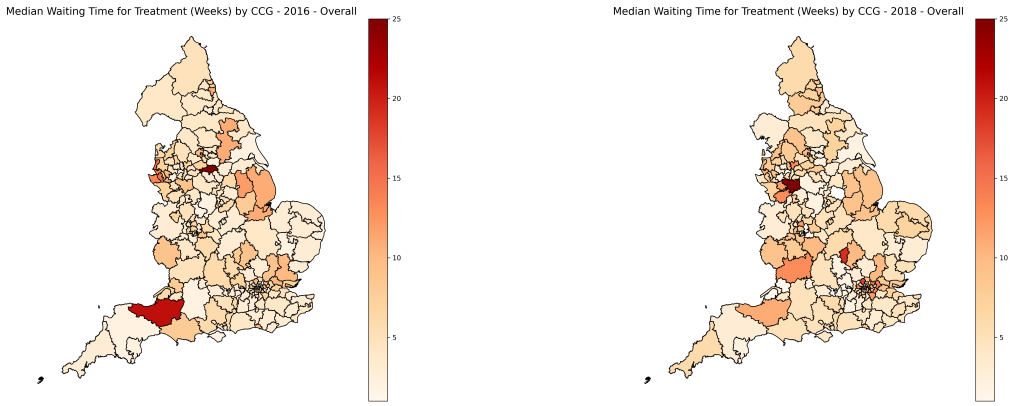
In 2018, the IAPT programme served about 17% of the community prevalence of depression and anxiety disorders. As a result, there was more demand for psychological therapies than there was supply. This oversubscription of patients to the programme creates exogenous variation in waiting times between initial assessment and first clinical session. This variation in waiting times differs across services (and, thereby, local areas) *and* over time depending on supply-side constraints, in particular shortages of trained therapists (due to staff recruitment and turnover or therapists undergoing training), and certain demand-side characteristics, in particular local clusters of mental ill health and regional differences in the balance between high-intensity and low-intensity treatment provision. Figure 1

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.

[16]Unfortunately, we do not have data on therapists, as data on patients cannot be linked to data on therapists.

shows substantial differences in median waiting times across areas and between 2016 and 2018.

Figure 1: Median Waiting Times in Weeks for Treatment by Clinical Commissioning Groups (CCGs)



(a) 2016 (b) 2018

This oversubscription of patients to the programme and resulting exogenous variation in waiting times informs our identification strategy, which we describe in Section 4.1, where we also provide more details on the variation in waiting times. In addition, we provide empirical evidence that waiting times are not systematically related to the severity of depression or anxiety symptoms at the start or end of treatment, the number of sessions, or treatment duration, neither across nor in any of the different intensity provisions of the IAPT programme.

2.2 Earlier Evaluations

Earlier evaluations of the IAPT programme provide only correlational evidence based on the comparison of patients' states before and after treatment. The first empirical study by Clark et al. (2009) evaluated two demonstration sites. The authors found a recovery rate of about 56%, which was largely maintained in a follow-up about ten months later.[17] 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 et al., 2018).[18]

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

[18] See Delgadillo et al. (2018) for an area-level analysis.

Another stream of evidence supporting the effectiveness of the IAPT programme comes from small-scale, short-run RCTs, testing new therapeutic approaches[19] or isolated components of the overall system.[20] 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 demonstrates that adding an IAPT-style psychological treatment in primary care is more (cost-)effective than treatment-as-usual (Cano-Vindel et al., 2022).

3 Data

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 all times, even if patients finish their therapy early. This is a useful design to avoid missing endline data, which could lead to an overestimation of the effectiveness of treatment. We define a course of treatment as including the initial assessment 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.[21]

The IAPT protocol requires patients to complete the same clinically validated measures of depression and anxiety in each session (including the initial assessment). 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.[22] Therapists then review these mea-

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

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

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

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

sures 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 valuable 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 first clinical session without any actual treatment occurring.

Our dataset consists of the universe of patients ever treated, entering the programme during the 2016 to 2018 period.[23] 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 associated local areas (e.g. local deprivation) from the Office for National Statistics (ONS) in the UK.

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 et al., 2001).[24] 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), likewise a routine instrument for measuring anxious affect and worry (Spitzer et al., 2006).[25] 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

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

[24]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.

[25]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.

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.[26]

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 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 additional outcomes, we also look at PHQ-9 and GAD-7 scores separately and at a mental health index, which is an average of both standardised scores. Note that PHQ-9 and GAD-7 as routine instruments in clinical practice are not considered to be prone to manipulation; any bunching observed in the distributions of these measures is due to different cut-offs for treatment.

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:

[26] 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.

mental ill health) overall as well as in different domains of life, including work, home management, social and private leisure, and close relationships.[27] 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. Just like our outcomes on mental health, data from the *Work and Social Adjustment Scale* and on the self-reported employment of patients are collected by the IAPT programme at the start of each session. Appendix Table A.I shows summary statistics of our outcomes.

Covariates. Patients' psychological-therapy characteristics include their referral type (whether they were referred by their GP or via self-referral), the time between referral and initial assessment in weeks, treatment mode (in person or online), whether they where prescribed additional medication (e.g. antidepressants), their initial diagnosis (depression and/or anxiety, including its type), and their treatment intensity (low or high-intensity treatment, and whether they changed their intensity during the course of treatment). Patients' individual characteristics include 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.

To capture supply-side constraints of the programme, the characteristics of services include the local number of staff, number of patients, and funding per patient. To capture demand-side characteristics, the socio-economic characteristics of associated local areas include the local unemployment rate and median wage as well as local deprivation (an index of multiple deprivation and sub-indices for deprivation in the areas of income, employment, education, health, crime, housing, and the environment). Appendix Table A.II shows summary statistics of our covariates. We routinely control for service and associated local-area characteristics throughout our regressions to capture supply and demand for therapies across areas and over

[27] 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 symptoms, 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."

time. Our results are robust to omitting any or all of these controls.

Estimation Sample. Our raw sample includes all patients who started treatment between April 2016 and December 2018. We focus on this 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 when estimating its causal policy effects. We remove courses of treatment that started in 2019 to not include patients that started in 2019 but did not finish by the time the Covid-19 pandemic disrupted data collection in early 2020. Our estimation sample therefore covers all data obtained from NHS Digital for the period in which the outcome monitoring system was fully operational; no data were dropped post-hoc.

We restrict this sample to attended sessions with non-missing values for both PHQ-9 and GAD-7 (recall that these are available for 98% of our sample). Moreover, we limit ourselves to patients who were at caseness prior to treatment, i.e. those who meet the clinical threshold for a mental health condition according to their PHQ-9 or GAD-7 scores at initial assessment.[28] The IAPT programme was launched precisely to serve these patients, making them its primary focus. Finally, we limit ourselves to patients who completed at least three sessions (the initial assessment and at least two subsequent clinical sessions), a requirement of our research design. We primarily study the effect of the full course of treatment, but in Section 5.2, we also look at the relative impact and value added of separate sessions cumulatively over the course of treatment. Our estimation sample includes 1,246,792 patients who attended, on average, 7.7 sessions (standard deviation of 4.1).[29]

[28] In special circumstances, therapists might accept individuals who do not meet treatment thresholds based on mental health scores if clinical judgment suggests the need for intervention. We do not include these patients in our sample, as they would qualify as recovered from the start and inflate the programme's effect.

[29] 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.

4 Empirical Strategy

4.1 Identification

Our aim is to estimate the causal effect of being treated within the IAPT programme. We use the potential outcomes framework by Rubin (1974), where 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.

Consider patient i who was assessed at time t and the duration of the (potential) treatment is w. For the moment, for the purpose of illustrating the main idea with lighter notation, take w as fixed and suppose we only consider a subset of the data for these patients. Let t_1 and t_2 respectively denote t and $t + w$. We introduce the following variables: D_i is a treatment dummy that takes value one for the treated; $Y_{it_j}(0)$ is the outcome for patient i at time t_j if they were to *not* receive treatment; $Y_{it_j}(1)$ is the outcome for patient i at time t_j if they were to receive treatment; and X_i is a vector of observed characteristics associated with patient i.

Our parameters of interest are ATT and CATT (conditional ATT) that we denote respectively by θ and $\theta(X_i)$. They are formally defined as:

$$\begin{aligned} \theta := &\quad E\left[Y_{it_2}(1) - Y_{it_2}(0) \,|D_i = 1\right], \\ \theta(X_i) := &\quad E\left[Y_{it_2}(1) - Y_{it_2}(0) \,|D_i = 1, X_i\right]. \end{aligned}$$

ATT and CATT are not identified without further assumptions since we only observe $Y_{it_j} := D_i Y_{it_j}(1) + (1 - D_i) Y_{it_j}(0)$, but never both $Y_{it_j}(1)$ and $Y_{it_j}(0)$. The identifying assumptions we make below are standard in the econometrics literature on difference-in-differences models when two time periods are available (e.g. see J. Roth et al. (2023)). In what follows, it is convenient to define $\Delta Y_i := Y_{it_2} - Y_{it_1}$ and $\Delta Y_i(d) := Y_{it_2}(d) - Y_{it_1}(d)$ for $d = 0, 1$. We assume the following assumptions hold throughout:

Assumption 1: Parallel trends. For all i**,**

$$E\left[\Delta Y_i(0) \,|D_i = 1, X_i\right] = E\left[\Delta Y_i(0) \,|D_i = 0, X_i\right] \text{ almost surely.}$$

Assumption 2: No anticipatory effects. For all i**,**

$$E\left[Y_{it_1}(0) \,|D_i = 1, X_i\right] = E\left[Y_{it_1}(1) \,|D_i = 1, X_i\right] \text{ almost surely.}$$

In our context, Assumption 1 states that the expected natural recovery for patients in the treatment and control group are the same without the IAPT programme. Assumption 2 states that the expected initial outcome, prior to 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 difference-in-differences in terms of observables, namely:

$$\theta = E\left[\Delta Y_i | D_i = 1\right] - E\left[\Delta Y_i | D_i = 0\right], \tag{1}$$

$$\theta\left(X_i\right) = E\left[\Delta Y_i | D_i = 1, X_i\right] - E\left[\Delta Y_i | D_i = 0, X_i\right]. \tag{2}$$

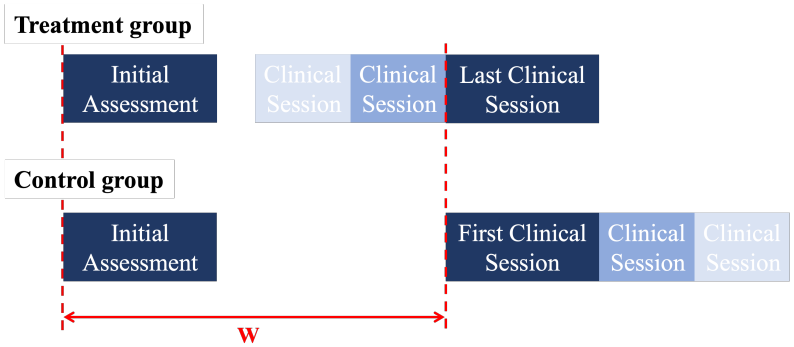
In Appendix B, we provide a proof that ATT and CATT can be written in terms of the distribution of observables, along with more detailed discussions.

We analyse our data through the lens of a two-period model, which is justified under the assumption that $\{(\Delta Y_i, D_i, X_i)\}_{i=1}^{n}$ is a random sample that, in turn, imposes the stable unit treatment value assumption and stationarity of the data generating process. This framework permits t_1 and t_2, hence $t_2 - t_1$, to vary across patients. Indeed, it is worth emphasising that our patients enter the programme at different times (and that we have no available pre-treatment data beyond the initial assessment and no available post-treatment data beyond the last session), so that our data are not suitable to be studied under a multi-period, cohort-wide adoption (i.e. a staggered design), 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 for identification, which creates exogenous variation in waiting times between initial assessment and the first clinical session across services (and, thereby, local areas) *and* over time. 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 (treatment group) with the change in mental health outcomes for patients between their initial assessment and their first clinical session (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 2 illustrates our research design.

Figure 2: Difference-in-Differences Design – Waitlist-Based Quasi-Randomisation



Note: Own illustration.

Given that X_i includes psychological-therapy, individual, service and associated local-area characteristics, as well as service and time fixed effects, we ensure that Assumptions 1 and 2 reasonably hold. Note: Assumption 1 is weaker than assuming that treatment assignment in our quasi-experiment is random conditional on X_i.

Patients are assigned to the treatment or control group based on their waiting times: the treatment indicator equals one if a patient's waiting time falls below a specified threshold, and zero otherwise. In our baseline model, we define this threshold using the median waiting time — ranging from 22 to 41 days, depending on whether patients receive low- or high-intensity treatment. Our results are robust to using alternative thresholds.

In what follows, we provide more details on the variation in waiting times and discuss potential residual selection.

4.1.1 Waiting Times

Our identification strategy exploits oversubscription of patients and resulting exogenous variation in waiting times between initial assessment and first clinical session across services (and, thereby, local areas) *and* over time. This variation depends on supply-side constraints, in particular shortages of trained therapists (due to staff recruitment and turnover or therapists undergoing training) and certain demand-side characteristics, in particular local clusters of mental ill health and regional dif-

ferences in the balance between high-intensity and low-intensity treatment provision. Appendix Section C documents in detail this variation across areas and over time. Appendix Figure C.I shows histograms of waiting times for our entire estimation sample and for individual years, Figure C.II histograms for different treatment intensities across all years. Table C.I presents relevant summary statistics. Figure C.III shows a heat map of median waiting times, overall and by treatment intensity, across services for all years. Figures C.IV to C.VI then replicate Figure C.III for each individual year.

Appendix E shows that waiting times are not systematically related to the severity of depression or anxiety symptoms at the start (e.g. PHQ-9 of 15.7 and GAD-7 of 14.3 in the 25th percentile of waiting time versus PHQ-9 of 15.9 and GAD-7 of 14.5 in the 90th) or end of treatment (e.g. PHQ-9 of 8.8 and GAD-7 of 7.9 versus PHQ-9 of 9.1 and GAD-7 of 8.2), the number of sessions (e.g. 7.8 versus 7.7 sessions), or treatment duration (e.g. 14.5 versus 13 weeks), neither across nor in any of the different intensity provisions of the IAPT programme. In Section 5.2, we show that, instead of using the median, using the 25th, 75th, and 90th percentile of waiting time to allocate patients into treatment and control group yields very similar results.[30]

Next, there may be concern that waiting times are correlated with certain service and associated local-area characteristics, which, in turn, may be correlated with patient outcomes, for example local deprivation. Section 4.1.2 shows that service and associated local-area characteristics are well-balanced between our treatment and control group, using the 50th percentile of waiting time as a default threshold.[31] As discussed, we routinely control for service and associated local-area characteristics, as well as service fixed effects, throughout our regressions, and our results are robust to omitting these controls.

Finally, there may be concern that the mental burden of long waiting time itself could have a negative impact on natural recovery, which, if true, can induce a false positive on the effectiveness of treatment. We argue that this is unlikely to be the case here, as waiting is expected by all patients. Criticisms of waiting times in

[30]This also suggests that the intensity of treatment *within* the treatment group is similar regardless of waiting time, supporting the notion of a stable unit treatment.

[31]The characteristics of services include the number of staff, number of registered patients, allocations per registered patient, unemployment rate, and median wage. The socio-economic characteristics of associated local areas include an index of multiple deprivation as well as sub-indices for deprivation by income, employment, education, health, crime, barriers to housing and services, and living environment.

the NHS have long been well-publicised, so having to wait is common knowledge. Moreover, the legal requirement to treat all patients fairly based on a strict first-come, first-serve protocol and associated waiting times are announced at initial assessment. Empirically, Appendix Figure C.VII plots our main outcomes — reliable recovery, improvement, and deterioration — as raw data for different waiting times. These figures show that outcomes are stable across waiting times and, arguably, even exhibit a minor natural recovery. That is, waitlisted patients are, if anything, more likely to improve than to deteriorate. Hence, positive estimates of treatment effects are due to therapy being beneficial, rather than from waiting being detrimental.

A potential concern may be that some patients experience longer waiting times because they missed scheduled appointments, which might correlate with slower natural recovery. Note, however, that under NHS England's IAPT waiting-time guidance, the referral-to-treatment clock does not stop when a patient cancels or misses an appointment. Hence, missed appointments do not mechanically inflate measured waiting times.

4.1.2 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 subsequent initial assessment, therapists allocate patients to either low- or high-intensity treatment, in each of which they are processed. As mentioned, this allocation is done following a strict first-come, first-served protocol, a legal requirement based on fairness principles, and is rigorously followed through.[32] In line with this, we observe only a weak, insignificant correlation between waiting time and either PHQ-9 or GAD-7 score.[33] Recall that we routinely control for pre-treatment mental health by including our mental health index, an average of both standardised PHQ-9 and GAD-7 scores, as well as the time lapsed between referral

[32] If present, prioritisation would lead to a lower-bound estimate. Note that, given a general shortage of therapists, higher-need patients are not systematically sent to more experienced therapists (within each treatment intensity), which would result in an upper bound treatment effectiveness estimate. As discussed in Section 2.1, therapists receive a standardised training with dedicated national curricula. To the extent that the initial assessment itself has therapeutic value, this does not bias our results as it is balanced between groups.

[33] $r = 0.017$ for PHQ-9, $r = 0.016$ for GAD-7.

and initial assessment in all our models.

To provide a more direct test of the exogeneity of waiting times, we regress waiting time on our mental health index, pre-treatment, controlling for psychological-therapy, individual, service and associated local-area characteristics, as well as service and time fixed effects. This is equivalent to asking whether baseline severity predicts waiting time conditional on the full set of controls used in our main specification. Appendix Table E.XIII shows that a full one standard deviation higher mental health index, pre-treatment, is associated with starting treatment *less than one day later*, on average. Although statistically significant due to the large sample size, this relationship is economically negligible. The overall R^2 of the model is 0.219, but this is driven almost entirely by service and time fixed effects and local-area characteristics that capture structural differences in capacity across areas and over time – precisely the supply- and demand-side variation that our identification strategy exploits. Baseline severity contributes negligibly to the explained variation in waiting times. Table E.XIII shows similar results by treatment intensity, confirming that the pattern holds within each stream of the programme's stepped-care model.

Next, Appendix Table D.I shows balancing properties of covariates — psychological-therapy, individual, service and associated local-area characteristics — between our treatment and control group, which uses the 50th percentile of waiting time as a default threshold. Following Imbens and Rubin (2015), we calculate four scale-free overlap measures: normalised differences (which, unlike simple differences in means, 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). Note that we routinely control for treatment mode in all our models. 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. Our covariates are, therefore, well balanced between groups.

Two differences between treatment and control group warrant further discussion. First, the treatment group is more likely to receive face-to-face treatment (35% versus 21%), whereas the control group is more likely to be treated via telephone (76% versus 61%). This pattern is consistent with the mechanics of the programme design: patients with shorter waiting times (treatment group) are more likely to have started treatment earlier in our observation period, when face-to-face delivery was more prevalent, whereas patients with longer waiting times (control group) are more likely to have started their first clinical session at a point when telephone-based delivery had become more established in the programme. As noted, we routinely control for treatment mode in all our models (our results are robust to omitting this control). Second, patients in the treatment group are slightly more likely to receive statutory sick pay at baseline (8.4% versus 7.1%) and slightly more likely to be employed as opposed to long-term sick (89.4% versus 86.7%). While their normalised differences remain well below the 0.25 threshold, they suggest small compositional differences in employment-related characteristics. We control for self-reported employment status throughout our regressions (our results are again robust to omitting this control). To the extent that any residual imbalance remains, it would likely bias our estimates downward, as patients on statutory sick pay or long-term sick leave tend to have more complex conditions and, as we show in Section 5.3, respond less favourably to treatment.

We note that non-response rates for certain demographic characteristics are high in both groups: for example, gender is not recorded for about 26% of patients, ethnicity for about 29%, religion for about 43%, and sexual orientation for about 40% (Appendix Table D.I). Non-response rates are slightly higher in the treatment than in the control group, though normalised differences between both groups remain well below 0.25 in all cases. These patterns likely reflect variation in administrative data collection practices across services and over time, rather than patient-level decisions to withhold information: the demographic items are collected as part of a standardised intake form, and completion depends in part on how consistently individual services enforce data entry. Importantly, our main outcomes – PHQ-9 and GAD-7 scores – are recorded for 98% of patients, as they are a mandatory part of the clinical protocol reviewed by therapists at the start of each session. Non-response in demographic covariates, therefore, does not affect the quality of our outcome data. Moreover, we include indicators for missing values in all relevant covariates throughout our regressions, ensuring that non-response does not

introduce bias through the exclusion of observations.

Finally, Appendix Table D.II shows balancing properties of outcomes — reliable recovery, improvement, and deterioration; PHQ-9 and GAD-7 scores; our mental health index; the *Work and Social Adjustment Scale*; and self-reported employment — between our treatment and control group at the start of different sessions. As seen, neither at initial assessment nor at the start of the first or last clinical session does any of the normalised differences exceed the recommended threshold of 0.25 (Imbens & Wooldridge, 2009). There is little evidence for an unusual dispersion of outcomes between groups at any point in time either. Patients in treatment and control 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 after waiting.

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 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 H, we establish bounds around our treatment effect estimates by imputing outcomes under various scenarios. We find that, even under the most extreme assumptions such that all those dropping out of the treatment group would experience deterioration and all those dropping out of the control group would experience recovery, our estimated treatment effects for both reliable recovery and reliable improvement remain significant and positive. Estimates under these assumptions are approximately half the magnitude of our baseline results. The programme continues to significantly reduce the likelihood of reliable deterioration, except in the most extreme scenario.

Similarly, individuals in our 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. It is worth noting that the IAPT programme is run by the NHS, which is the monopolist provider of state-funded healthcare in England. It was launched precisely because patients had few other treatment options available.

4.2 Estimation

4.2.1 Average Treatment Effects

In Section 4.1, we only consider patients at time t that have w weeks as the duration of or waiting time for treatment. We now combine observations for different t's and w's and update our notation by letting $\Delta Y_i := D_i \Delta Y_i^{tr} + (1 - D_i) \Delta Y_i^c$, with $\Delta Y_i^{tr} := Y_{it_i+W_i}(1) - Y_{it_i}(1)$ and $\Delta Y_i^c := Y_{it_i+W_i}(0) - Y_{it_i}(0)$. That is, ΔY_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, ΔY_i^{tr}; and the change between initial assessment and the first clinical session if i belongs to our control group, ΔY_i^c, cf. Figure 2. Due to the importance of the duration of or waiting time for treatment in our model, we treat this separately from other covariates and denote it by W_i. W_i denotes the duration of treatment or waiting time respectively for a patient in the treatment or control group. D_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 22 and 41 days, depending on the intensity of treatment.[34] X_i contains all other relevant observables including psychological-therapy, individual, service and associated local-area characteristics, type of service and time.

We assume $\{(\Delta Y_i, D_i, W_i, X_i)\}_{i=1}^n$ to be i.i.d. and expand the conditioning set in Assumptions 1 and 2 to include W_i. While we can identify heterogeneous treatments under this assumption, in this subsection we focus on a model with homogeneous treatment and suppose that the following holds:

$$E[\Delta Y_i | D_i, W_i, X_i] = \beta_0 + \beta_1 D_i + \beta_2^\top \widetilde{W}_i + \beta_3^\top \widetilde{X}_{it_i} + \mu_{r_i} + \nu_{t_i}. \tag{3}$$

Then, β_1 represents the ATT. Here, W_i is represented by a vector, $\widetilde{W}_i$, of binary variables indicating the weeks in which a patient completed either treatment or waiting so that $\beta_2^\top \widetilde{W}_i$ captures the, possibly non-linear, effect of natural recovery for patient i. X_i contains some variables that can vary with time for different patients, and it is decomposed into $\widetilde{X}_{it_i}$, which contains psychological-therapy, individual, service and associated local-area characteristics, and μ_{r_i} and ν_{t_i} are, respectively, service (i.e. 135 CCGs) and time fixed effects (i.e. day-of-week, month, and year).

[34] The median threshold is 27 days for low and 22 days for high-intensity treatment, 35 days for stepped-up courses, 41 days for stepped-down courses, and 32 days if the treatment intensity is undefined (due to multiple changes).

Including both service and time fixed effects implies that we are looking at variation across services (and, thereby, geographical areas) *and* over time. We also routinely control for waiting time and time lapsed between referral and initial assessment in weeks as well as for pre-treatment mental health (in form of our mental health index) throughout.

We estimate the following model:

$$\Delta Y_i = \beta_0 + \beta_1 D_i + \beta_2^\top \widetilde{W}_i + \beta_3^\top \widetilde{X}_{it_i} + \mu_{r_i} + \nu_{t_i} + u_i. \tag{4}$$

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 across services and over time. We estimate treatment effects in Equation 4 using OLS with robust standard errors clustered at the service level.[35]

4.2.2 Heterogeneous Treatment Effects

Under Equation 3, the treatment effect is assumed to be the same for all types of patients. To 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 correlational 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.

[35] Given that ΔY_i is discrete for out main outcomes, in Section 5.2, we provide the results of logit model as a robustness check.

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 $\{(\Delta Y_i, D_i, W_i, Q_i)\}_{i=1}^n$. To facilitate matching, we dichotomise the covariates identified in earlier correlational studies as related to heterogeneity in treatment outcomes. Each combination of these dichotomised covariates defines a patient type, represented by Q_i, which replaces X_i as the type indicator for each patient. Our CATT is then indexed by (w, q), which corresponds to a particular treatment/waiting time duration and patient type. In this case, as alluded to earlier, our CATT is identified and can be written for each (w, q) as (cf. Equation 2),

$$\theta(w, q) := E[\Delta Y_i^{tr} | W_i = w, Q_i = q] - E[\Delta Y_i^{c} | W_i = w, Q_i = q]. \quad (5)$$

Since (W_i, Q_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 patients whose $W_i = w$ and $Q_i = q$. We only include sub-populations that have a sufficient number of observations for both treatment and control group.[36] 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 ensures 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 & Zinn, 1990). Conveniently, however, the nonparametric estimator just described is numerically equivalent to the OLS estimator of $\{\theta(w, q)\}$ from the following model:

$$\begin{aligned} \Delta Y_i &= \sum_{w,q} \beta(w, q) \times \mathbf{1}\{Q_i = q, W_i = w\} \\ &\quad + \sum_{w,q} \theta(w, q) \times \mathbf{1}\{Q_i = q, W_i = w\} \times D_i + u_i. \end{aligned} \quad (6)$$

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

where $\mathbf{1}\{Q_i = q, W_i = w\}$ 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 $\{\theta(w, q)\}$. For example, one can simply test the homogeneity hypothesis on the CATTs, where the null hypothesis states that all CATTs are equal, using a Wald test.

ATT with data-driven sources of heterogeneity. To further explore heterogeneities without constraining the analysis to a set of pre-selected sources, we use the *generalised random forest* (Athey et al., 2019).

The algorithm recursively splits the sample into two bins, with each bin subsequently split further. This process continues iteratively, creating a tree-like structure. Somewhat similar to our nonparametric approach, the bins share the same realisations of covariates. The difference is that the partitioning into bins does not rely on the researcher's choice of covariates but is done in a data-driven way to maximise heterogeneity in within-bin treatment effect estimates across bins. This partitioning process is repeated multiple times, generating several trees. The individual treatment effect estimates from these trees are then averaged to reduce variance, ultimately providing individual-level CATT estimates.[37]

To take it to a more familiar context, a forest can be thought of as a nearest-neighbour method, in that it performs the estimation using a weighted average of observations in the "neighbourhood". However, in contrast to classical methods, the neighbourhood is defined in a flexible data-driven way. By treating the forest as an adaptive nearest-neighbour estimator, Athey et al. (2019) show that the estimates of the generalised random forest are consistent and asymptotically normal.[38]

[37] 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).

[38] See Athey et al. (2019) and Wager and Athey (2018) for a 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 report the results for models that control for psychological-therapy, individual, service and associated local-area characteristics, as well as service and time fixed effects, which are our preferred 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. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

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.[39]

[39] Overall, 53% of patients reliably recover at the end of the treatment, 74% reliably improve, and 5% reliably deteriorate.

The latter suggests, in particular, 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 regardless of whether we include covariates or not.

To gauge effect sizes, the control-group means for the change in outcomes, which capture natural recovery or deterioration during the waiting period, are approximately 9% for reliable recovery, 36% for reliable improvement, and 14% for reliable deterioration. Our estimated treatment effects are, therefore, large relative to these baselines: treatment increases reliable recovery roughly fivefold and reliable improvement by about two-thirds, while reducing deterioration by more than half. Note that the estimated treatment effect on reliable recovery is larger than on reliable improvement, even though recovery is the stricter criterion. This is because the natural recovery rate in the control group (9%) is much lower than the natural improvement rate (36%), leaving substantially more room for treatment to increase recovery relative to its baseline.

Treatment Intensity. 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 those in the high-intensity treatment, and Panel C for those who are stepped up from initially low to then high intensity. The full results, which include smaller streams (patients who are stepped down from initially high to then low intensity or patients for whom the intensity was not recorded), are presented in Appendix Table F.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 trained therapists to different treatment intensities results in an appropriate patient-therapy fit.

The similarity of recovery across treatment intensities may tempt one to think that different intensities are redundant if these lead to similar outcomes. Note, however, that patients in different treatment intensities have different therapeutic needs. Appendix Table F.II 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. As seen, patients in the high-intensity treatment show much stronger symptom 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 therapists (re-)allocate patients to suitable treatments, if needed, and that different treatment intensities cater to different needs, which is also reflected in differences in underlying therapies and mechanisms, as outlined in Section 2.

Work and Social Life Outcomes. 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). 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.

Appendix Table F.III 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 F.IV shows our average treatment effects on employment as a result of treatment.[40] As seen, being treated within the IAPT programme has, overall, 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 like-

[40]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 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 a value-added model. Note that all of our previous results continue to hold when using this alternative model.

lihood 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 evidence for small, positive short-term impacts on employment of patients who undergo psychological therapy.

Note that some IAPT services have introduced *Employment Advisors* who provide employment support alongside psychological therapy (cf. Thew et al., 2023). During our observation period (2016 to 2018), such provision was not yet widespread and limited to a small number of pilot sites only. Our employment effects should, therefore, be interpreted primarily as a consequence of improved mental health through psychological therapy rather than of direct employment support. To the extent that some patients in our sample did receive employment advice in pilot sites, this would be captured by our service fixed effects.

5.2 Robustness Checks

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

Selection on Outcome and Session Value Added. So far, we restricted our estimation sample to patients who completed a course of treatment, consisting of initial assessment and at least two subsequent clinical sessions. There could be concern that the timing of the last clinical session may be endogenous, i.e. therapists may discard patients after they have reached a particular threshold of recovery, and patients then leave the programme). To check for selection on the outcome, we exploit our session-by-session patient-level outcome data to redefine the completion of treatment, to not pertain to the last but to the penultimate clinical session or even to the one before. Arguably, the latter two sessions should not be susceptible to selection on the outcome. Appendix Table G.I shows that, for both redefinitions of treatment completion, we continue to detect strong, positive impacts of treatment on mental health. Naturally, impacts are somewhat reduced, as we omit clinical sessions with therapeutical contents, which are particularly relevant for courses of treatment with a lower number of total sessions. Note that the number of observations drops because we lose particularly short courses of treatment. A related concern regarding selection on the outcome could be that therapists may

switch patients from, say, low to high intensity because their health may deteriorate. When grouping together low-intensity and step-up as well as high-intensity and step-down, here too we continue to detect strong, positive impacts of treatment on mental health (Appendix Table G.II).

To further address potential selection on the outcomes, i.e. the total number of sessions or treatment duration, we estimate a model with the same selection on outcomes for treatment and control groups. Appendix Table G.III additionally controls for the total number of clinical sessions and the total duration of treatment in weeks. As seen, our results remain robust. Note that we routinely control for a large set of pre-treatment characteristics including patients' psychological-therapy characteristics, including their referral type (whether they were referred by their GP or via self-referral), the time between referral and initial assessment in weeks, treatment mode (in person or online), whether they where prescribed additional medication (e.g. antidepressants), their initial diagnosis (depression and/or anxiety, including its type), and their treatment intensity (low or high-intensity treatment, and whether they changed their intensity during the course of treatment).

Our session-by-session patient-level outcome data also allow us to look at the relative impact and value added of separate sessions cumulatively over the course of treatment. Appendix Figure G.I 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. In each case, the control group is restricted to patients who have the same number of total sessions. We observe that the relative session value added is lower for patients who have a higher total number of sessions. 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. Moreover, 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. A possible concern may be that this pattern reflects selective dropout, that is, patients leaving treatment after a positive shock, rather than genuine ther-

apeutic gains. However, several features of our analysis and the institutional setting mitigate this concern. First, Figures G.I to G.III condition on the total number of sessions, meaning that they compare patients with the *same* planned course length. The pattern, therefore, cannot be driven by compositional changes due to dropout across session bins. Second, within the IAPT programme, therapists and patients jointly agree on a course of treatment at the outset, based on the initial diagnosis and severity of symptoms, with national guidelines specifying recommended session numbers for different conditions. Hence, course length is largely pre-determined by clinical need rather than by within-treatment recovery shocks. Patients with more severe or complex conditions are planned for longer courses, which explains why the per-session value added is lower for patients with more total sessions: these patients require more sessions to achieve the same recovery threshold. Third, as shown in Appendix Table G.I, we continue to detect strong treatment effects when redefining treatment completion to the penultimate or antepenultimate session, which should not be susceptible to endogenous termination.

Appendix Figures G.II and G.III replicate Figure G.I for reliable improvement and deterioration, showing a similar pattern. Note that, as the spacing of sessions is about two weeks and hence very similar to the recall period of both PHQ-9 and GAD-7 (see next paragraph), we do not expect that our results on session value added are confounded by double-counting of symptoms.

Session Spacing. Typically, patients are meant to have one session per week, though the median number of weeks between sessions is 1.6 (mean of 2.0 and standard deviation of 1.6), depending on patients' availability. Hence, the actually observed session spacing is very similar to the recall period of both PHQ-9 and GAD-7, which asks patients about symptoms in the past two weeks. Appendix Table G.IV shows that our results are robust to additionally controlling for the average number of weeks between sessions.

Appendix Table G.V then makes full use of our session-by-session patient-level outcome data to look at the spacing of sessions over the course of treatment, by re-estimating our average treatment effects by percentile of the number of weeks between sessions. We differentiate the lower 25th percentile (session spacing of 1.1 weeks) from the upper 25th (2.4 weeks) and the upper 10th percentile (3.5 weeks). As seen, reliable recovery and improvement are slightly higher the lower the number of weeks between sessions. Although the variation in session spacing is rather

small, a caveat of this analysis is that session spacing may be partly endogenous, for example if reasons for rescheduling sessions are correlated with aspects of mental health.

Repeat Enrolment. Repeat enrolment may be a sign of poor mental health amongst dropouts. We observe that, in total, 187,148 patients (about 15%) enrol more than once in the IAPT programme. To check whether waitlisted patients in our default control group drop out and present again later, Appendix Table G.VI regresses the likelihood to enrol more than once in the programme on the weeks on the waitlist amongst control-group patients, with and without controls. As seen, the weeks on the waitlist have negligible predictive power for repeat enrolment. As patients who repeatedly enrol may be special in other ways too, Appendix Table G.VII excludes them altogether from our analysis. As seen, our results remain similar to before.

Selective Attrition. We address potential concerns about selective attrition in detail in Appendix Section H, where we show that the programme remains highly effective for reliable improvement and reliable recovery even under extreme assumptions on the outcomes of patients who discontinue treatment or drop out, though the magnitude of the effects varies under different assumptions. The programme also continues to significantly reduce the likelihood of reliable deterioration, except in the scenario with the most extreme assumptions, i.e. that all dropped-out respondents who would have been assigned to the treatment group experience deterioration and all those dropping out of the control group experience recovery.

Note that, in our bounding analysis, we impute waiting times for patients who dropped out based on the average waiting time for their assigned treatment intensity at their service in the month of assessment. While a richer imputation model incorporating additional covariates such as baseline severity or demographic characteristics could reduce noise in the imputed treatment assignment, this would not affect the bounds themselves, which are constructed by assuming extreme outcomes (full recovery or full deterioration) for all dropouts regardless of their imputed treatment status. Our approach is, therefore, conservative by design.

Other Robustness Checks. Our results are robust to different definitions of treatment and control group when varying treatment and corresponding waiting time durations. In Section 4.1, we have shown that waiting times are not system-

atically related to the number of sessions, treatment duration, or the severity of depression or anxiety symptoms at the start or end of treatment. So far, we used the median waiting time as a cut-off to define treatment and control group. Appendix Table G.VIII now uses, instead of the 50th percentile of waiting time, the 25th, 75th, and 90th percentile, respectively, to allocate patients into treatment and control group. The estimates remain similar. This stability is also consistent with our earlier finding that waiting times are largely orthogonal to baseline severity, treatment intensity, and other patient characteristics (see Section 4.1.1 Appendix Table E.XIII): varying the threshold primarily shifts the relative sizes of treatment and control group rather than meaningfully changing their composition.

Our results are also robust to different models, samples, and outcomes. Appendix Table G.IX 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 & Bretterville-Jensen, 2022), whereas Column 3 focuses only on patients who have depression and anxiety disorders, the main target population of the IAPT programme and vast majority. 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. In all cases, our results remain robust.

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 based on the 50th percentile of waiting time.

Results with pre-selected sources of heterogeneity. 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, service funding, and area deprivation.[41] [42] Figure 3 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.[43]

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.

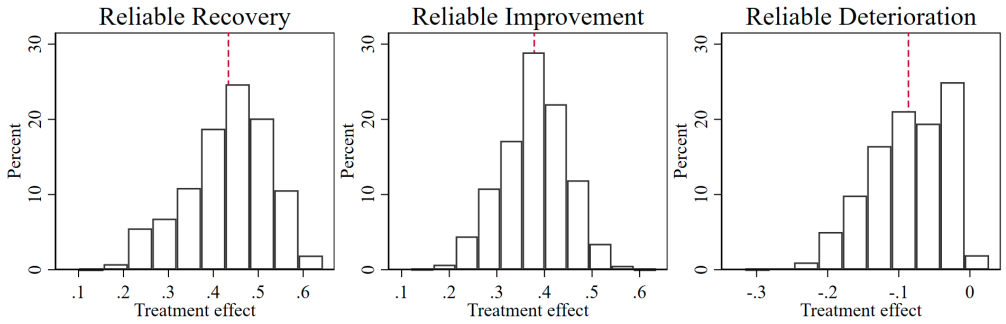


Figure 3: Conditional Average Treatment Effects – Matching Approach

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, and service and associated local area characteristics. The estimates are weighted by the number of treatment-group observations in each sub-population.

[41] 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; Delgadillo et al. (2016): a binary indicator for area deprivation above the median.

[42] 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 I.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%.

[43] 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 our ML analysis, 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.

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

$$\begin{aligned}\Delta Y_i = \beta_0 + \beta_1 D_i + \sum_q \beta_q Q_i + \sum_w \beta_w W_i \\ + \sum_q \gamma_q Q_i D_i + \sum_w \gamma_w W_i D_i + u_i,\end{aligned} \tag{7}$$

where, to assess how the effect of the treatment differs for different sub-populations, the treatment dummy, D_i, is interacted with the psychological-therapy and individual as well as service and associated local-area characteristics, Q_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 I.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.

Table 3: Heterogeneous Treatment Effects on Mental Health: Pre-Defined Sources

	Reliable Recovery (1)	Reliable Improvement (2)	Reliable Deterioration (3)
Course Intensity:			
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	0.021***	0.026***	-0.010***
Median * Treated	(0.002)	(0.002)	(0.001)
Religion:			
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:			
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)
Number of Individuals	947,547	947,547	947,547
R Squared	0.26	0.16	0.05

Note: Linear probability models. Binary dependent variables. Robust standard errors clustered at service level in parentheses. Omitted categories are Low Intensity, Christian, White British. The full results are presented in Appendix Table I.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.[44]

[44]Note that we do not include gender or socio-economic status as separate sources of heterogeneity in Table 5.3, as gender is not recorded for about 26% of patients, making subgroup estimates less reliable. Moreover, socio-economic status is not directly observed at the individual level but captured only indirectly through employment status and area deprivation, both of which are included in our analysis and emerge as significant sources of heterogeneity.

For area characteristics, patients in more deprived areas are less likely to reliably recover, which is in line with the findings of Delgadillo et al. (2016). The effect size is similar to having a long-term health condition. However, these patients are moderately 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. 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, whilst funding of the services is positively related.

Results with data-driven sources of heterogeneity. Figure 4 presents the histograms of our heterogeneous treatment effect estimates produced by the generalised random forest described in Section 4.2.2.[45] The vertical dashed line again 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 improvement are bounded away from zero, whilst reliable deterioration is not.

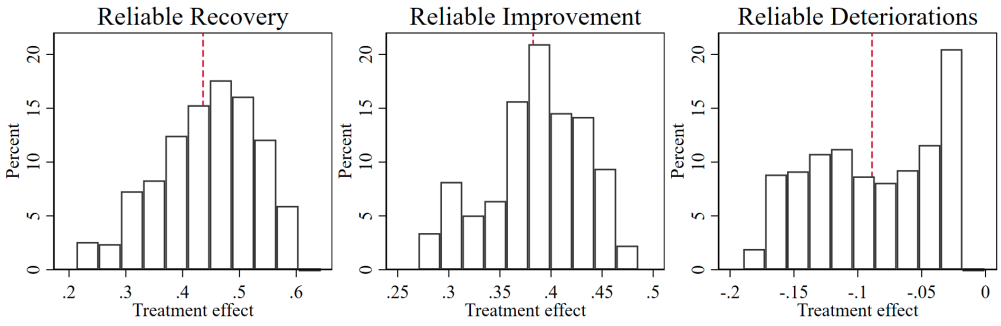


Figure 4: Conditional Average Treatment Effects – Generalised Random Forest

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, individual, service and associated

[45]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.

local-area characteristics in sup-populations formed by quartiles of the estimated treatment effect distribution. The first quartile includes individuals whose estimated treatment effects were in the bottom 25% of all estimated individual treatment effects, the second to fourth quartiles are formed accordingly. Appendix Tables I.III, I.IV, and I.V report the results for all covariates. Here, we discuss covariates that show substantial difference across quartiles.

First, the results of the data-driven approach support the findings from the previous section. Patients who are less likely to recover tend to exhibit more severe symptoms at the start of treatment. They are also more likely to live in deprived areas, attend larger services as indicated by the number of patients, or have their gender, ethnicity, sexual orientation, or disability status not recorded. These patterns largely hold for reliable improvement, where, in addition, patients who are less likely to improve attend services that, on average, have lower funding. 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 treatment and to live in more deprived areas.

Second, the ML algorithm provides two new insights: patients who recover less are more likely to be unemployed at the start of treatment, whilst patients who recover more are more likely to self-refer. To study these sources in a more systematic way, we estimate a modification of Equation 4, where the treatment dummy is now interacted with each of the two newly identified sources. Table 4 reports the results for reliable recovery, reliable improvement, and reliable deterioration.

Table 4: Heterogeneous Treatment Effects on Mental Health: Sources Identified in the ML Algorithm

	Reliable Recovery (1)	Reliable Improvement (2)	Reliable Deterioration (3)
Unemployed vs. Employed			
Treated	0.468***	0.387***	-0.085***
	(0.004)	(0.004)	(0.001)
Unemployed	-0.012***	-0.083***	0.029***
	(0.001)	(0.003)	(0.002)
Unemployed * Treated	-0.133***	-0.042***	0.009***
	(0.004)	(0.004)	(0.002)

Number of Individuals	828,356	828,356	828,356
R Squared	0.30	0.19	0.06
Self Referral vs. Non-Self Referral			
Treated	0.404***	0.373***	-0.089***
	(0.005)	(0.004)	(0.002)
Self Referral	0.016***	0.043***	-0.022***
	(0.006)	(0.006)	(0.004)
Self Referral * Treated	0.038***	0.006	0.007***
	(0.005)	(0.004)	(0.002)
Number of Individuals	1,246,792	1,246,792	1,246,792
R Squared	0.29	0.19	0.06
Therapy Controls	Yes	Yes	Yes
Individual Controls	Yes	Yes	Yes
Service Controls	Yes	Yes	Yes
Local-Area Controls	Yes	Yes	Yes
Service Fixed Effects	Yes	Yes	Yes
Time Fixed Effects	Yes	Yes	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.

Unemployment emerges as a significant source of heterogeneity for both patients awaiting treatment and those undergoing it, even after controlling for a rich set of covariates, including the severity of symptoms. The first panel of Table 4 presents the results of the comparison of employed and unemployed patients.[46] Unemployed patients are 1.2 percentage points less likely to recover naturally while on the waitlist. Additionally, they are 13.3 percentage points less likely to recover as a result of treatment, which represents 30% of the programme's average treatment effect. Unemployed individuals are also less likely to reliably improve and more likely to deteriorate, thought the magnitude of the estimate for the latter is very small so it is unlikely to be economically meaningful. Earlier studies generally agree that unemployment negatively affects mental health (Cygan-Rehm et al., 2017), highlighting the need for public policy to prioritise early prevention of mental health issues amongst the unemployed. We provide suggestive evidence that unemployed patients, on average, respond to treatment less favourably than their employed counterparts.

Several mechanisms may explain this finding. First, unemployed patients face

[46]The number of observations is lower than in other specifications because we exclude individuals with other employment statuses.

ongoing stressors, such as financial insecurity and social isolation, which persist throughout the course of treatment, potentially limiting the scope for CBT to bring about lasting impact. Second, unemployment may proxy for a broader set of disadvantages, including lower social support or more complex co-morbidities, which are not fully captured by baseline PHQ-9 and GAD-7 scores even after controlling for severity. Third, practical barriers associated with unemployment, such as lack of daily routine or transport costs, may reduce treatment engagement. These mechanisms are not mutually exclusive and suggest that combining psychological therapy with employment support could enhance treatment effectiveness for this group, an approach that some IAPT services have begun to pilot (cf. Thew et al., 2023).

Self-referral is an unusual and possibly controversial feature of the IAPT programme.[47] The possibility to self-referral contrasts with other healthcare in the UK as well as with healthcare provision in other countries, such as Denmark or the Netherlands. Hence, the IAPT dataset provides a unique opportunity to study how these patients respond to treatment.[48] Self-referred patients constitute 71.5% of our sample. As noted, they sought care considerably earlier than patients referred via other pathways: on average, 364 days after the onset of symptoms compared to 461 days.[49] Self-referral is well balanced between our treatment and control group, with rates of 72.1% and 70.9%, respectively (Appendix Table D.I). We routinely control for referral type in all our models.

The second panel of Table 4 presents the results of the comparison of treatment effects for self-referred patients with those who accessed treatment via other pathways. We find that patients who self-referred are 1.6 percentage points more likely to recover while on the waitlist and 3.8 percentage point more likely to recover as the result of treatment, which represents 8% of the average treatment effect. Self-referred patients are also more likely to reliably improve and less likely to reliably

[47] See Brown et al. (2010) for a discussion of advantages, e.g. improved access, and disadvantages, e.g. system overload due to relatively minor cases, of self-referrals.

[48] Anecdotal evidence suggests that some patients who self-referred to the programme did so at the recommendation of their GP. Since participation in the programme demands a certain level of commitment, clinicians might use self-referral as a way to ensure that patients are more likely to remain engaged if they choose to join independently. In our data, we cannot differentiate between those who were informally referred by their GP and those who discovered the programme on their own, so we analyse these groups together. 71.5% of all patients in our sample self-referred. All patients are assessed in the same way, regardless of the referral type.

[49] We observe a self-reported date of symptom onset for approximately a third of the total sample. We remove observations where the date of onset was recorded after the referral date.

deteriorate while on the waitlist. The effects of being treated on the probability of reliable improvement and deterioration are similar for self-referred and non-self-referred patients. The coefficient on the difference for self-referred participants in the latter case is statistically significant, but its magnitude is very small, making it not economically meaningful. These findings underline the importance of using causal methods for treatment effect estimation: more favourable outcomes would appear in a correlational analysis from differences in natural recovery rates rather than a difference in treatment effect estimates. We need to caveat that, within this framework, we cannot investigate what drives the differences in the outcomes of patients who self-referred compared to those referred through other pathways. It is possible that they have higher motivation to improve their mental health or differ in other non-observable characteristics.

In sum, this is suggestive evidence that the option of self-referrals improves access to mental healthcare.

5.4 Cost-Benefit Calculation

We perform a simple and conservative cost-benefit calculation. In doing so, we compare being treated within the IAPT programme to business-as-usual prior to IAPT, which in most cases was no treatment at all.[50] Note that we routinely control for medication usage in treatment and control, as pharmacology could be a complement (or substitute) to IAPT.[51]

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 G.IX). 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).[52] UK Government values 1.0 QALYs at £70,000 (in

[50] Recall that the IAPT programme was launched precisely because there was a lack of treatment options for mild to moderate common mental health problems in the UK. Besides IAPT, there were (and are) community mental health services in the UK, but these are targeting primarily severe cases. To our knowledge, there exists no systematic evaluation of these services.

[51] We do not find that being treated within the IAPT programme reduces medication usage, if used (results available upon request).

[52] 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/.

2019 prices) (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 only 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.[53]

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 the mental health of patients, our primary outcome and unit of analysis. Unfortunately, we do not have data on secondary outcomes and on partners. However, it is well-documented in the economics literature 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 example spillovers on significant others (such as partners, children, or the wider community). Reichman et al. (2015) show that being out of 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 treasury in form of additional tax income and

[53] An alternative way to look at benefits is to use *Wellbeing-Years (WELLBYs)* (Frijters & Krekel, 2021; Frijters et al., 2020). Noting that an increase in the EQ-5D-5L index of 0.03 points translates into an increase in WELLBYs of 0.11 (using a conversion factor of 1 EQ-5D-5L = 3.79 WELLBYs, see Frijters and Krekel, 2021 Table 3A.4), and that 1.0 WELLBYs is valued by HM Treasury at £13,000 (Treasury, 2021), we obtain monetised benefits of (((0.00 + 0.11) / 2) * 2 months + (0.11 * 0.6) * 10 months) / 12 months + 0.11 * 0.6 * 2 years * £13,000 = £2,550 per patient over a three-year period. This yields net benefits of £2,550 - £680 = £1,870 per patient three years after the end of treatment, or a benefit-cost ratio of 3.8.

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 et al., 1999; Clark & Layard, 2014). In a Norwegian RCT study of an IAPT-style intervention, Smith et al. (2024) find that income (and hence taxes) increase significantly two to three years after the end of therapy. This has also been found in a Spanish context (Munoz-Navarro et al., 2024).[54] This has led some authors to argue that public savings in terms of taxes and benefits alone would turn net public costs negative, making the programme pay for itself (Layard, 2016). As we observe patients only from start to end of therapy, we remain conservative and focus only on benefits in terms of mental health, which by themselves already suggest that the programme is worth it.

6 Discussion and Conclusion

Mental ill health deeply affects individuals, their families, and society, while also posing a substantial economic challenge. Yet, it is often relegated to the sidelines of healthcare priorities, overshadowed by physical health issues. This does not have to be the case, as there are now successful examples of evidence-based programmes that address mental health needs.

This paper is the first to estimate the causal effects of a nationwide mental health service at a scale that well represents the English patient population. We use data on all patients who started treatment in the IAPT programme between April 2016 and December 2018 and exploit oversubscription and resulting exogenous variations in waiting times across services and over time for identification. Our empirical strategy can be used to evaluate the effectiveness of other public services too, in contexts where demand for services exceeds supply, leading to variations in waiting times. Our dataset is of an exceptionally high quality in terms of its outcome completeness and worldwide unique in terms of its session-by-session outcome monitoring, thus offering a data generating process with precisely the variation we need to identify causal effects based on exogenous variation in waiting times, including specific features such as session value added. It is exceedingly difficult, if not impossible, to find a better dataset than ours to answer our research

[54] Serena (2025), however, finds no long-term labour market effects of extending health insurance coverage of psychotherapy in Denmark up to seven years after treatment. The author looks at prime-working-age patients between 18 and 37 years with mild-to-moderate symptoms.

question. Despite its uniqueness, our dataset and method allow us to provide generalisable insights into the functioning of nationwide-scaled mental health services, not only in England but also in other countries, in particular those with IAPT-style programmes such as in Australia, Canada (Ontario), Lithuania, Norway, Spain, or Sweden.

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. We found that the programme provides significant mental health benefits. In particular, 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 or those residing in deprived areas.

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 cost-benefit calculation 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 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.

References

Angelucci, M., & Bennett, D. (2024). The Economic Impact of Depression Treatment in India: Evidence from Community-Based Provision of Pharmacotherapy. *American Economic Review*, *114*(1), 169–198.

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

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148–1178.

Baigent, M., Smith, D., Battersby, M., Lawn, S., Redpath, P., & McCoy, A. (2023). The Australian version of IAPT: Clinical outcomes of the multi-site cohort study of NewAccess. *Journal of Mental Health*, *32*(1), 341–350.

Banerjee, S., Chatterji, P., & Lahiri, K. (2017). Effects of Psychiatric Disorders on Labor Market Outcomes: A Latent Variable Approach Using Multiple Clinical Indicators: Psychiatric Disorders and Labor Market Outcomes. *Health Economics*, *26*(2), 184–205.

Baranov, V., Bhalotra, S., Biroli, P., & Maselko, J. (2020). Maternal Depression, Women's Empowerment, and Parental Investment: Evidence from a Randomized Controlled Trial. *American Economic Review*, *110*(3), 824–859.

Barker, N., Bryan, G., Karlan, D., Ofori-Atta, A., & Udry, C. (2022). Cognitive Behavioral Therapy among Ghana's Rural Poor Is Effective Regardless of Baseline Mental Distress. *American Economic Review: Insights*, *4*(4), 527–545.

Batistich, M. K., Evans, W. N., Giles, T., & Margolit-Chan, R. (2024). *Therapy to Reduce Violence and Improve Institutional Safety During Incarceration* (NBER Discussion Paper No. 33147). National Bureau of Economic Research.

Beam, E. A., & Quimbo, S. (2023). The Impact of Short-Term Employment for Low-Income Youth: Experimental Evidence from the Philippines. *Review of Economics and Statistics*, *105*(6), 1379–1393.

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

Berger, M. C., & Black, D. A. (1992). Child Care Subsidies, Quality of Care, and the Labor Supply of Low-Income, Single Mothers. *Review of Economics and Statistics*, *74*(4), 635–642.

Bhalotra, S., Daysal, N. M., & Trandafir, M. (2025). *Antidepressant Treatment in Childhood* (IFS Working Paper No. 25/38). Institute for Fiscal Studies.

Biasi, B., Dahl, M. S., & Moser, P. (2026). Career Effects of Mental Health: Evidence from an Innovation in Treating Bipolar Disorder. *Journal of Political Economy: Microeconomics*, *forthcoming*.

Blattman, C., Jamison, J. C., & Sheridan, M. (2017). Reducing Crime and Violence: Experimental Evidence from Cognitive Behavioral Therapy in Liberia. *American Economic Review*, *107*(4), 1165–1206.

Bossuroy, T., Goldstein, M., Karimou, B., Karlan, D., Kazianga, H., Parenté, W., Thomas, C. C., Udry, C., Vaillant, J., & Wright, K. A. (2022). Tackling psychosocial and capital constraints to alleviate poverty. *Nature*, *605*, 291–297.

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

Brown, J. S., Boardman, J., Whittinger, N., & Ashworth, M. (2010). Can a self-referral system help improve access to psychological treatments? *British Journal of General Practice*, *60*(574), 365–371.

Cano-Vindel, A., Ruiz-Rodríguez, P., Moriana, J. A., Medrano, L. A., González-Blanch, C., Aguirre, E., & Muñoz-Navarro, R. (2022). Improving Access to Psychological Therapies in Spain: From IAPT to PsicAP. *Psicothema*, (34.1), 18–24.

Chatterji, P., Alegria, M., & Takeuchi, D. (2011). Psychiatric disorders and labor market outcomes: Evidence from the National Comorbidity Survey-Replication. *Journal of Health Economics*, *30*(5), 858–868.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), 1–68.

Chiles, J. A., Lambert, M. J., & Hatch, A. L. (1999). The Impact of Psychological Interventions on Medical Cost Offset: A Meta-analytic Review. *Clinical Psychology: Science and Practice*, *6*(2), 204–220.

Cho, H. J., Lavretsky, H., Olmstead, R., Levin, M., Oxman, M. N., & Irwin, M. R. (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), 442–451.

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

Clark, D. M., Canvin, L., Green, J., Layard, R., Pilling, S., & Janecka, M. (2018). Transparency about the outcomes of mental health services (IAPT approach): an analysis of public data. *Lancet*, *391*, 679–686.

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

Clark, D. M., & Layard, R. (2014). *Thrive: The Power of Psychological Therapy*. Penguin.

Clark, D. M., Wild, J., Warnock-Parkes, E., Sott, R., Grey, N., Thew, G., & Ehlers, A. (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), 5022–5032.

Connor, K. M., Davidson, J. R. T., Churchill, L. E., Sherwood, A., Weisler, R. H., & Foa, E. (2000). Psychometric properties of the Social Phobia Inventory (SPIN). *British Journal of Psychiatry*, *176*(4), 379–386.

Costantini, S. (2025). How Do Mental Health Treatment Delays Impact Long-Term Mortality? *American Economic Review*, *115*(5), 1672–1707.

Cronin, C. J., Forsstrom, M. P., & Papageorge, N. W. (2024). What Good Are Treatment Effects Without Treatment? Mental Health and the Reluctance to Use Talk Therapy. *Review of Economic Studies*, rdae061.

Cuddy, E., & Currie, J. (2026). Rules versus Discretion: Treatment of Mental Illness in US Adolescents. *Journal of Political Economy*, *134*(1), 478–522.

Cuijpers, P., Cristea, I., Karyotaki, E., Reijnders, M., & Huibers, M. J. H. (2016). How effective are cognitive behavior therapies for major depression and anxiety disorders? A meta-analytic update of the evidence. *World Psychiatry*, *15*(3), 245–258.

Cuijpers, P., Smit, F., Bohlmeijer, E., Hollon, S. D., & Andersson, G. (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), 173–178.

Cygan-Rehm, K., Kuehnle, D., & Oberfichtner, M. (2017). Bounding the causal effect of unemployment on mental health: Nonparametric evidence from four countries. *Health Economics*, *26*(12), 1844–1861.

Dague, L., DeLeire, T., & Leininger, L. (2017). The Effect of Public Insurance Coverage for Childless Adults on Labor Supply. *American Economic Journal: Economic Policy*, *9*(2), 124–154.

Delgadillo, J., Asaria, M., Ali, S., & Gilbody, S. (2018). On poverty, politics and psychology: the socioeconomic gradient of mental healthcare utilisation and outcomes. *British Journal of Psychiatry*, *209*(5), 429–430.

Delgadillo, J., Asaria, M., Ali, S., & Gilbody, S. (2016). On poverty, politics and psychology: The socioeconomic gradient of mental healthcare utilisation and outcomes. *British Journal of Psychiatry*, *209*(5), 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.

Dinerstein, M., Megalokonomou, R., & Yannelis, C. (2022). Human Capital Depreciation and Returns to Experience. *American Economic Review*, *112*(11), 3725–3762.

Ehlers, A., Wild, J., Warnock-Parkes, E., Grey, N., Murray, H., Kerr, A., Rozental, A., Thew, G., Janecka, M., Beierl, E. T., Tsiachristas, A., Perer-Salazar, R., Andersson, G., & Clark, D. M. (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), 608–622.

Fava, G. A., Ruini, C., Rafanelli, C., Finos, L., Conti, S., & Grandi, S. (2004). Six-Year Outcome of Cognitive Behavior Therapy for Prevention of Recurrent Depression. *American Journal of Psychiatry*, *161*(10), 1872–1876.

Finkelstein, A., Hendren, N., & Luttmer, E. F. P. (2019). The Value of Medicaid: Interpreting Results from the Oregon Health Insurance Experiment. *Journal of Political Economy*, *127*(6), 2836–2874.

Fletcher, J. (2009). All in the Family: Mental Health Spillover Effects between Working Spouses. *The B.E. Journal of Economic Analysis & Policy*, *9*(1).

Fletcher, J. M. (2010). Adolescent depression and educational attainment: Results using sibling fixed effects. *Health Economics*, *19*(7), 855–871.

Fonagy, P., Lemma, A., Target, M., O'Keefe, S., Constantinou, M. P., Wurman, T. V., Luyten, P., Allison, E., Roth, A., Cape, J., & Pilling, S. (2019). Dynamic interpersonal therapy for moderate to severe depression: a pilot randomized controlled and feasibility trial. *Psychological Medicine*, *50*(6), 1010–1019.

Frijters, P., Clark, A. E., Krekel, C., & Layard, R. (2020). A happy choice: wellbeing as the goal of government. *Behavioural Public Policy*, *4*(S2), 126–165.

Frijters, P., & Krekel, C. (2021). *A Handbook for Wellbeing Policy-Making*. Oxford University Press.

Frijters, P., Johnston, D. W., & Shields, M. A. (2014). The effect of mental health on employment: Evidence from Australian panel data. *Health Economics*, *23*(9), 1058–1071.

Frisch, R., & Waugh, F. V. (1933). Partial time regressions as compared with individual trends. *Econometrica*, *1*(4), 387–401. Retrieved 2023, from http://www.jstor.org/stable/1907330.

Furukawa, T. A., Levine, S. Z., Buntrock, C., Ebert, D. D., Gilbody, S., Brabyn, S., Kessler, D., Björkelund, C., Eriksson, M., Kleiboer, A., van Straten, A., Riper, H., Montero-Marin, J., Garcia-Campayo, J., Phillips, R., Schneider, J., Cuijpers, P., & Karyotaki, E. (2021). How can we estimate QALYs based on PHQ-9 scores? Equipercentile linking analysis of PHQ-9 and EQ-5D. *BMJ Mental Health*, *24*, 97–101.

Ghosal, S., Mani, A., Jana, S., Mitra, S., & Roy, S. (2022). Sex Workers, Stigma and Self-Image: Evidence from Kolkata Brothels. *Review of Economics and Statistics*, *104*(3), 431–448.

Gine, E., & Zinn, J. (1990). Bootstrapping general empirical measures. *The Annals of Probability*, *18*(2), 851–869. Retrieved 2023, from http://www.jstor.org/stable/2244320.

Gruber, J., Lordan, G., Pilling, S., Propper, C., & Saunders, R. (2022). The impact of mental health support for the chronically ill on hospital utilisation: Evidence from the UK. *Social Science & Medicine*, *294*, 114675.

Gyani, A., Shafran, R., Layard, R., & Clark, D. M. (2013). Enhancing recovery rates: Lessons from year one of IAPT. *Behaviour Research and Therapy*, *51*(9), 597–606.

Harvey, L. J., White, F. A., Hunt, C., & Abbott, M. (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*, 104408.

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

Heller, S. B., Shah, A. K., Guryan, J., Ludwig, J., Mullainathan, S., & Pollack, H. (2017). Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago. *Quarterly Journal of Economics*, *132*(1), 1–54.

Hoe, T. P. (2023). Does Hospital Crowding Matter? Evidence from Trauma and Orthopedics in England. *American Economic Journal: Economic Policy*, *14*(2), 231–236.

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

Imbens, G. W., & Wooldridge, J. M. (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, *47*(1), 5–86.

Jacob, B. A., Kapustin, M., & Ludwig, J. (2015). The Impact of Housing Assistance on Child Outcomes: Evidence from a Randomized Housing Lottery. *Quarterly Journal of Economics*, *130*(1), 465–506.

Jacob, B. A., & Ludwig, J. (2012). The Effects of Housing Assistance on Labor Supply: Evidence from a Voucher Lottery. *American Economic Review*, *102*(1), 272–304.

Johnsen, T. J., & Friborg, O. (2015). The effects of cognitive behavioral therapy as an anti-depressive treatment is falling: A meta-analysis. *Psychological Bulletin*, *141*(4), 747–768.

Kalin, N. H. (2020). The Critical Relationship Between Anxiety and Depression. *American Journal of Psychiatry*, *177*(5), 365–367.

Knapstad, M., Lervik, L. V., Sæther, S. M. M., Aarø, L. E., & Smith, O. R. F. (2020). Effectiveness of Prompt Mental Health Care, the Norwegian Version of Improving Access to Psychological Therapies: A Randomized Controlled Trial. *Psychotherapy and Psychosomatics*, *89*(2), 90–105.

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a Brief Depression Severity Measure. *Journal of General Internal Medicine*, *16*(9), 606–613.

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

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

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

Moller, N. P., Ryan, G., Rollings, J., & Barkham, M. (2019). The 2018 UK NHS Digital annual report on the Improving Access to Psychological Therapies programme: a brief commentary. *BMC Psychiatry*, *19*, 252.

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

Munoz-Navarro, R., Saunders, R., Buckman, J. E. J., Ruiz-Rodriguez, P., Gonzalez-Blanch, C., Medrano, L. A., Moriana, J. A., & Cano-Vindel, A. (2024). Investing in mental health: a path to economic growth through psychological therapies. *British Journal of Psychiatry*, *225*, 460–461.

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

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

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

NHS. (2019). NHS Long-Term Plan.

NHS. (2021a). Psychological Therapies, Annual Report on the Use of IAPT Services, 2020-21.

NHS. (2021b). What were clinical commissioning groups?

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

Richards, D., Enrique, A., Eilert, N., Franklin, M., Palacois, J., Duffy, D., Earley, C., Chapman, J., Jell, G., Sollesse, S., & Timulak, L. (2020). A pragmatic randomized waitlist-controlled effectiveness and cost-effectiveness trial of digital interventions for depression and anxiety. *npj Digital Medicine*, *3*, 85.

Richards, D., & Suckling, R. (2009). Improving access to psychological therapies: Phase IV prospective cohort study. *British Journal of Clinical Psychology*, *48*(4), 377–396.

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

Roth, A., & Fonagy, D. (2005). *What Works for Whom? A Critical Review of Psychotherapy Research*. Guilford Press.

Roth, J., Sant'Anna, P. H., Bilinski, A., & Poe, J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*, *235*(2), 2218–2244.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701.

Serena, B. L. (2025). The Causal Effect of Scaling up Access to Psychotherapy [_-eprint: https://direct.mit.edu/rest/article-pdf/doi/10.1162/REST.a.1679/2572288/rest.a.1679.pdf *The Review of Economics and Statistics*, 1–52.

Sevim, D., Baranov, V., Bhalotra, S., Maselko, J., & Biroli, P. (2023a). *Socioemotional Skills in Early Childhood: Evidence from a Maternal Psychosocial Intervention* (IZA Discussion Paper No. 15925). Institute of Labor Economics.

Sevim, D., Baranov, V., Bhalotra, S., Maselko, J., & Biroli, P. (2023b). *Trajectories of Early Childhood Skill Development and Maternal Mental Health* (Working Paper No. 1469). University of Warwick, Department of Economics.

Smith, O. R. F., Clark, D. M., Hensing, G., Layard, R., & Knapstad, M. (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., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097.

Strauss, C., Bibby-Jones, A.-M., Jones, F., Byford, S., Heslin, M., Parry, G., Barkham, M., Lea, L., Crane, R., de Visser, R., Arbon, A., Rosten, C., & Cavanagh, K. (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 (LIGHT-Mind) Randomized Clinical Trial. *JAMA Psychiatry*, *80*(5), 415–424.

Thew, G., Popa, A., Allsop, C., Crozier, E., Landsberg, J., & Sadler, S. (2023). The addition of employment support alongside psychological therapy enhances

the chance of recovery for clients most at risk of poor clinical outcomes. *Behavioural and Cognitive Psychotherapy*.

Toffolutti, V., Stuckler, D., McKee, M., Wolsey, I., Chapman, J., Pimm, T. J., Ryder, J., Salt, H., & Clark, D. M. (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), 224–233.

Treasury, H. (2021). Wellbeing Guidance for Appraisal: Supplementary Green Book Guidance.

Treasury, H. (2022). The Green Book.

Wager, S., & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.

Wakefield, S., Kellett, S., Simmonds-Buckley, M., Stockton, D., Bradbury, A., & Delgadillo, J. (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), 1–37.

WHO. (2022). *World mental health report: Transforming mental health for all*. World Health Organization. https://iris.who.int/bitstream/handle/10665/356115/9789240050860-eng.pdf?sequence=1.

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

Williams, J., & Bretterville-Jensen, A. L. (2022). *What's Another Day? The Effects of Wait Time for Substance Abuse Treatment on Health-Care Utilization, Employment and Crime* (IZA Discussion Paper No. 15083). IZA - Institute of Labor Economics.

Appendix

A Summary Statistics

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
Therapy Controls		
Mental Health Index (Pre-Treatment)	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
Individual Controls		
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
Service Controls		
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
Local-Area Controls		
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

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 $(\Delta Y_i, D_i, X_i)$.

Proof. Under Assumption 1, expanding out $\Delta Y_i(0)$ and re-arranging gives:

$$E[Y_{it_2}(0)|D_i = 1, X_i] = E[Y_{it_1}(0)|D_i = 1, X_i] + E[\Delta Y_i(0)|D_i = 0, X_i].$$

By Assumption 2, the first term on the right-hand-side of the equation above becomes $E[Y_{it_1}(1)|D_i = 1, X_i]$, so that $E[Y_{it_2}(0)|D_i = 1, X_i]$ is equal to $E[Y_{it_1}|D_i = 1, X_i] + E[Y_{it_2} - Y_{it_1}|D_i = 0, X_i]$. Subsequently, CATT is identified from the joint distribution of $(\Delta Y_i, D_i, X_i)$ since,

$$\theta(X_i) = E[\Delta Y_i|D_i = 1, X_i] - E[\Delta Y_i|D_i = 0, X_i].$$

Hence, ATT is also identified because, by the law of iterated expectation, $\theta = E[\theta(X_i)|D_i = 1]$.■

The proof strategy used in Proposition 1 is the conditional version of the one 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 $P(D_i|X_i)$ to be uniformly bounded away from 1 almost surely and $E[D_i] > 0$. 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,

$$\Delta Y_i = \sum_{w,q} [\beta(w, q) + \theta(w, q) \times D_i] \times \mathbf{1}\{Q_i = q, W_i = w\} + u_i,$$

which has the following matrix representation,

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

where $\mathbf{\Delta Y}$ is an $n \times 1$ vector of $\{\Delta Y_i\}_{i=1}^n$, $\iota(w, q)$ 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 $(W_i = w, Q_i = q)$ and $(D_i = 1, W_i = w, Q_i = q,)$, and $\mathbf{u}$ is a vector of $\{u_i\}_{i=1}^n$. 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 & Waugh, 1933), the OLS estimator from estimating (6) is the same as the OLS estimator obtained from estimating,

$$\Delta Y_i = \beta(w, q) + \theta(w, q) \times D_i + u_i,$$

when only observations of i's that correspond to $(W_i = w, Q_i = q)$ 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). This proves our claim. ■

C Summary Statistics on Waiting Times

Figure C.I: Histograms for Waiting Times in Weeks, All Treatments Intensities, All Years and by Year.

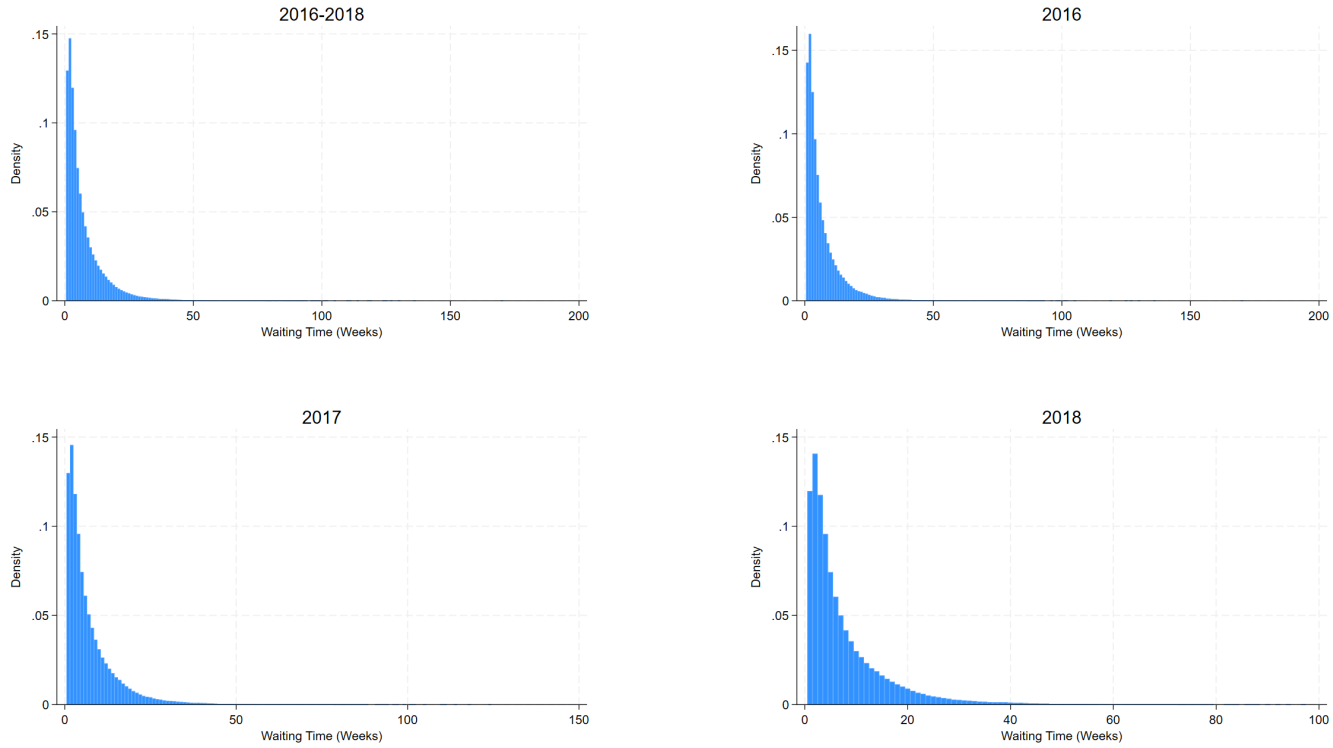
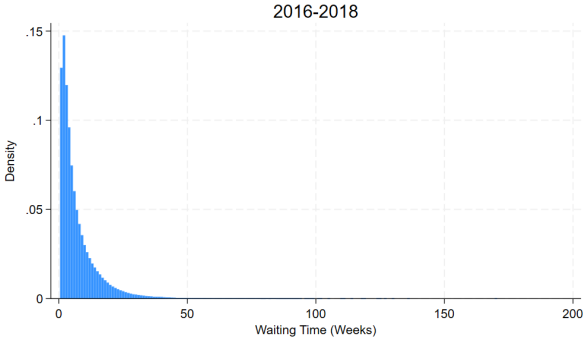
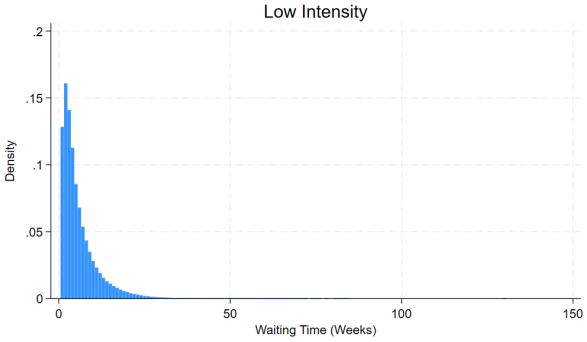


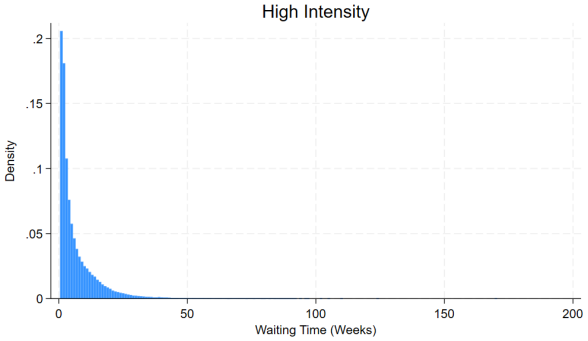
Figure C.II: Histograms for Waiting Times, by Treatments Intensities, All Years.



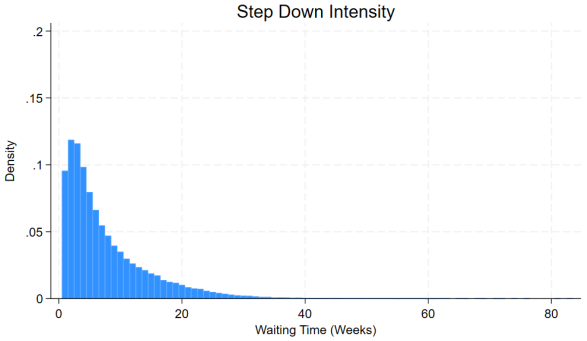
(a) All Intensities



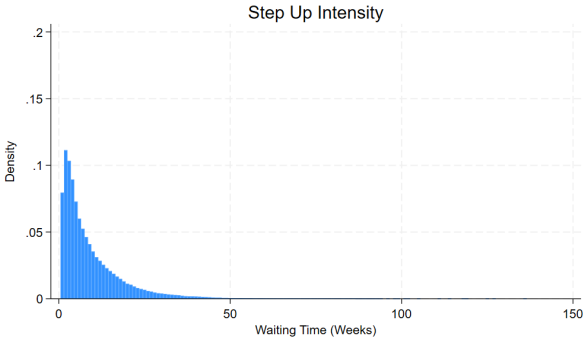
(b) Low Intensity



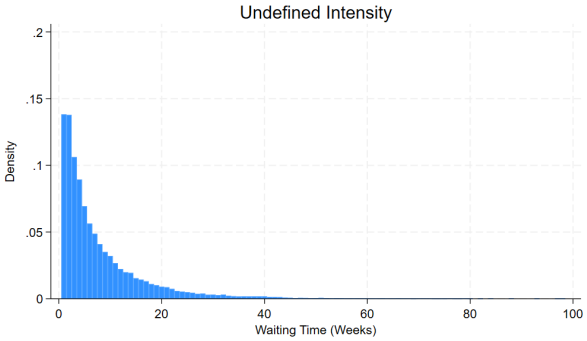
(c) High Intensity



(d) Step Down



(e) Step Up

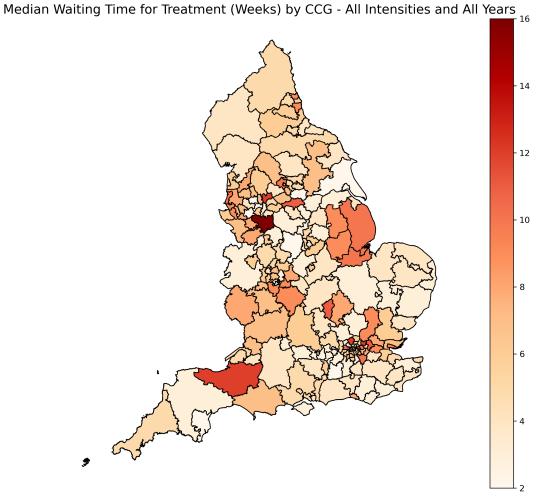


(f) Undefined

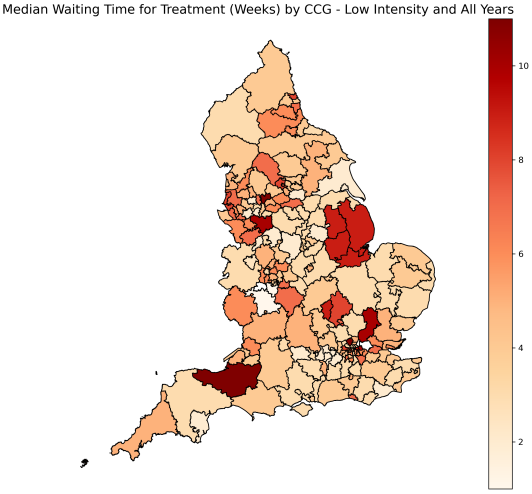
Table C.I: Summary Statistics for Waiting Time in Weeks, by Treatment Intensity

Variable	Mean	SD
Low Intensity	5.857	5.553
High Intensity	6.669	7.910
Step Down	7.780	7.195
Step Up	9.349	9.324
Undefined	8.057	8.898

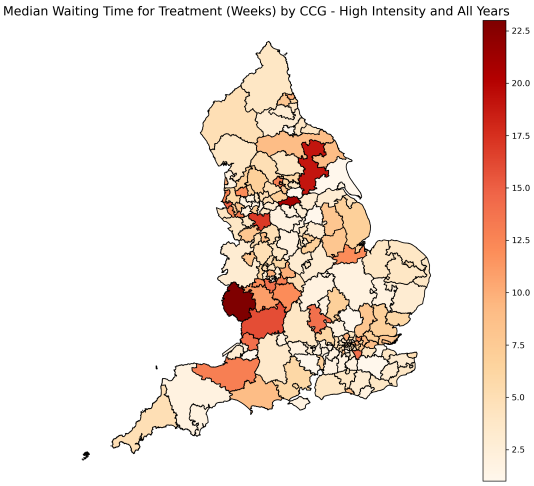
Figure C.III: Median Waiting Times in Weeks for Treatment by Clinical Commissioning Groups (CCGs) and Treatment Intensity, All Years



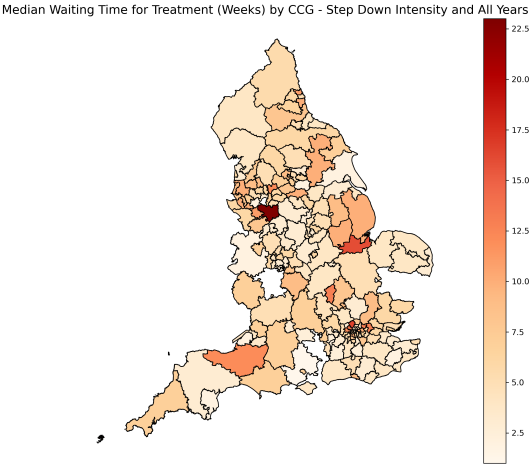
(a) All Intensities



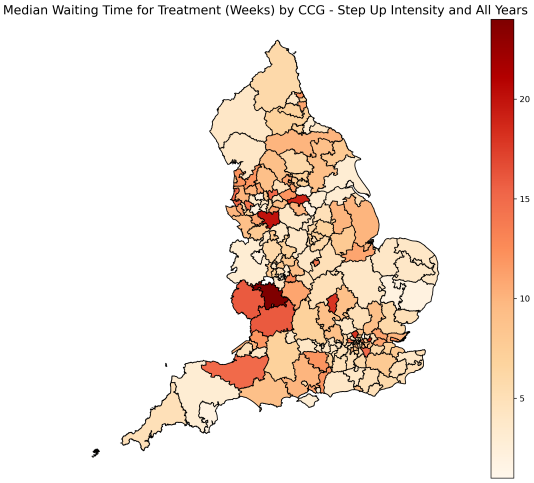
(b) Low Intensity



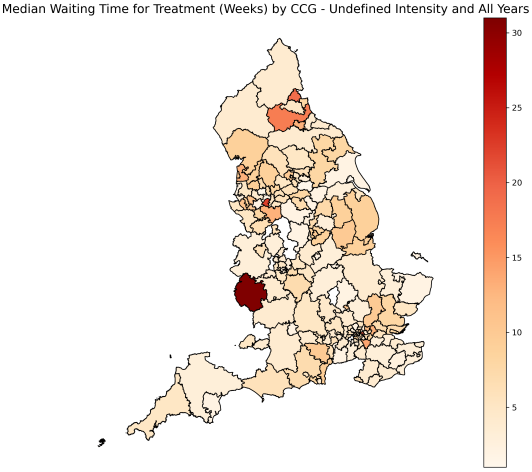
(c) High Intensity



(d) Step Down

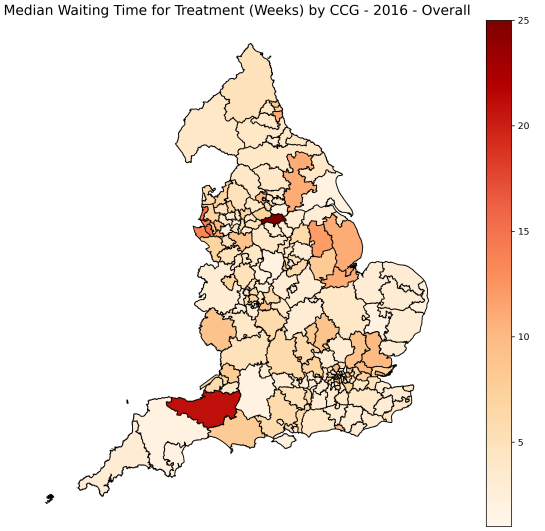


(e) Step Up

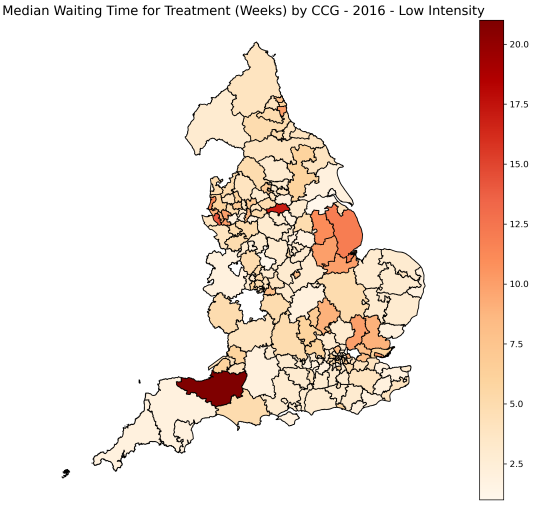


(f) Undefined

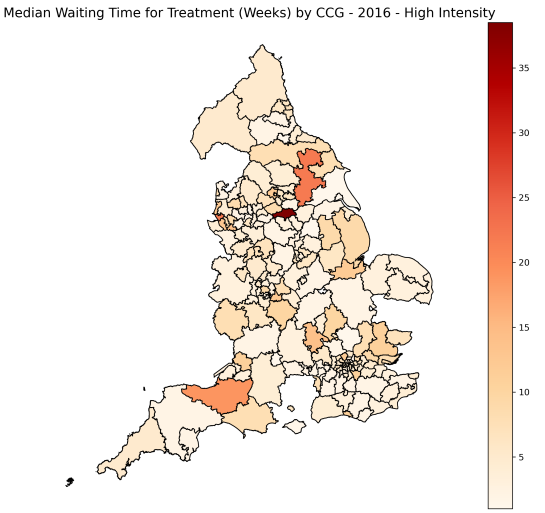
Figure C.IV: Median Waiting Times in Weeks for Treatment by Clinical Commissioning Groups (CCGs) and Treatment Intensity, 2016



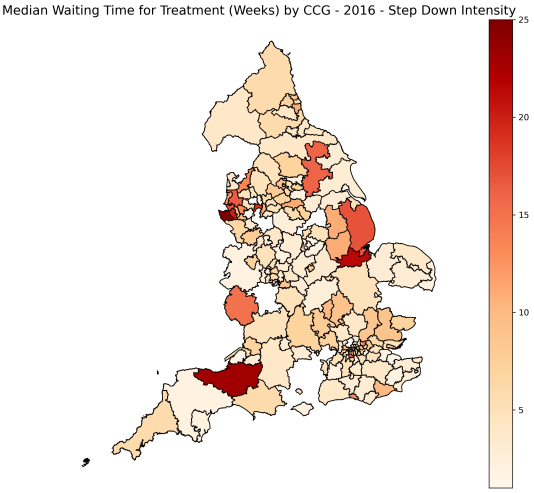
(a) All Intensities



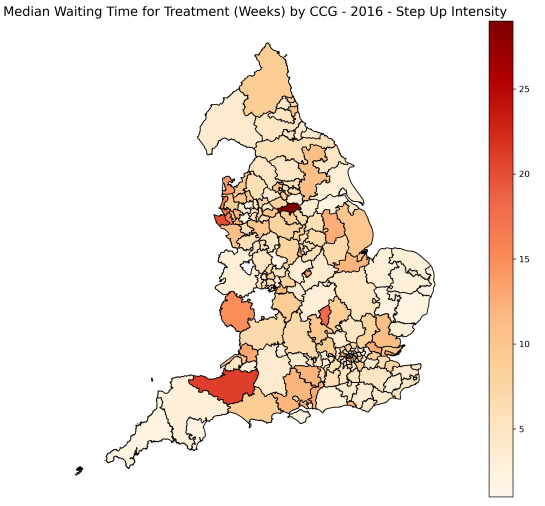
(b) Low Intensity



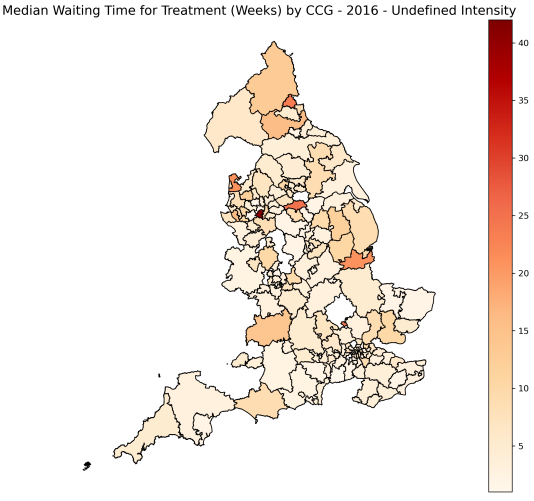
(c) High Intensity



(d) Step Down

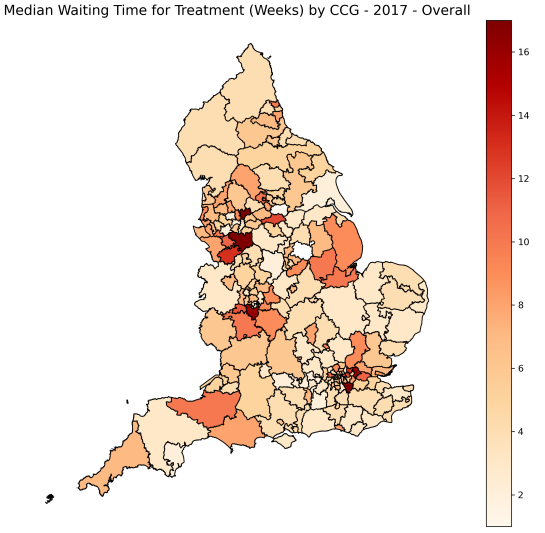


(e) Step Up

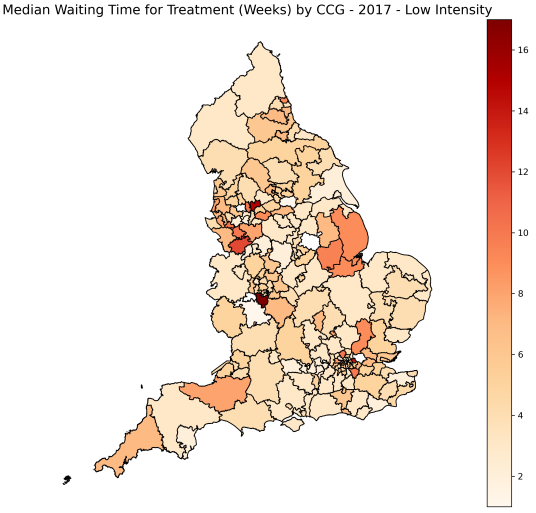


(f) Undefined

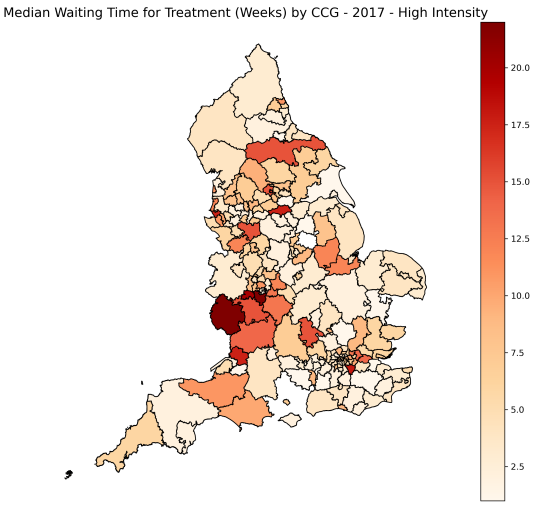
Figure C.V: Median Waiting Times in Weeks for Treatment by Clinical Commissioning Groups (CCGs) and Treatment Intensity, 2017



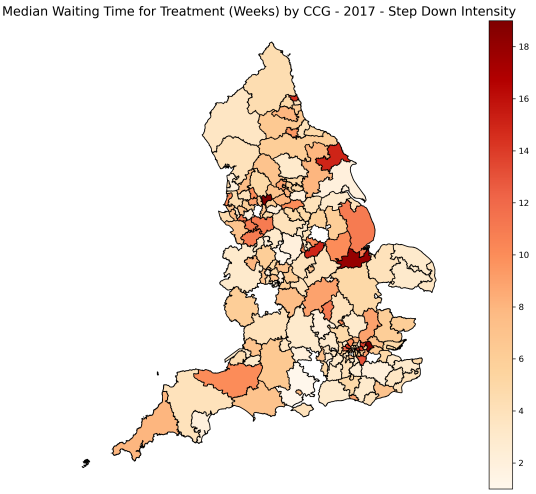
(a) All Intensities



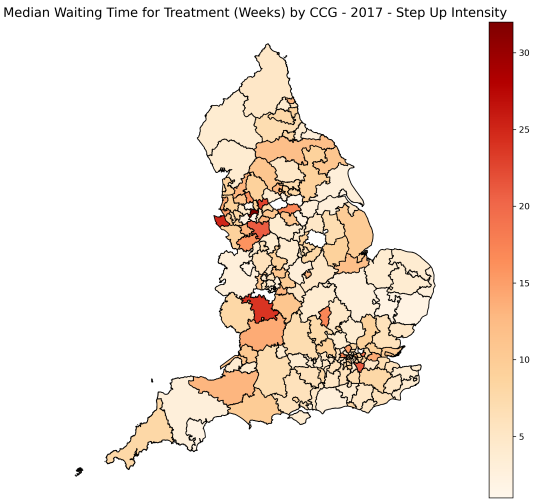
(b) Low Intensity



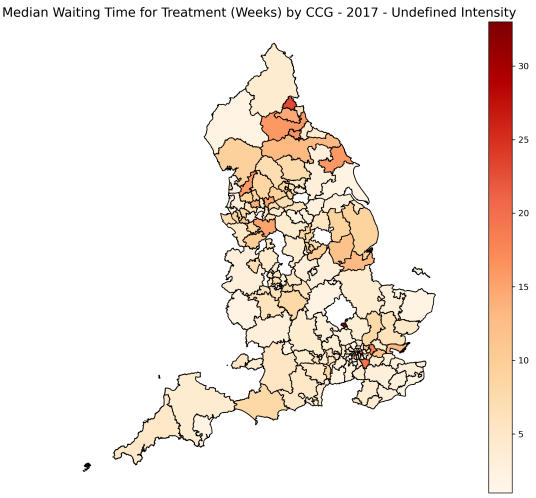
(c) High Intensity



(d) Step Down

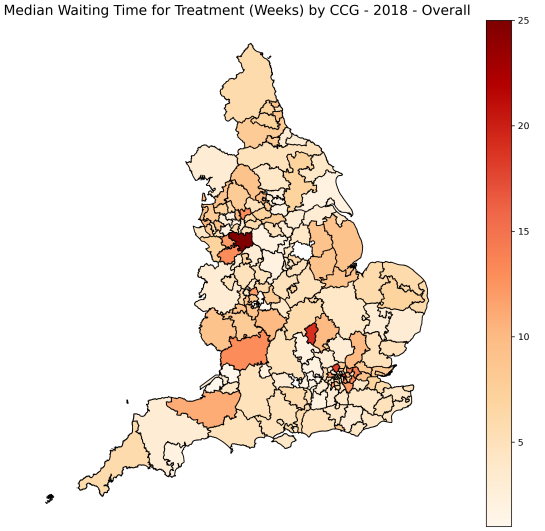


(e) Step Up

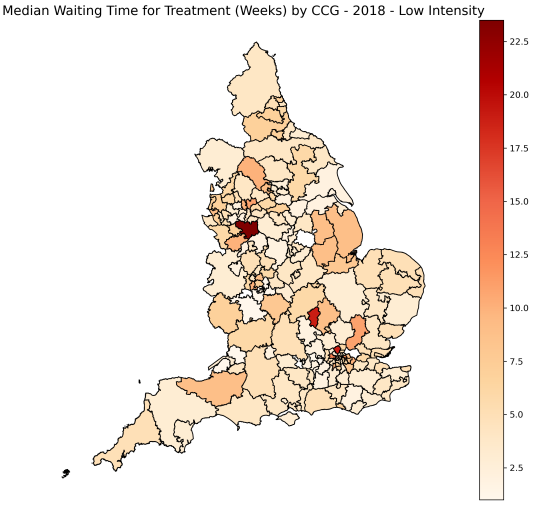


(f) Undefined

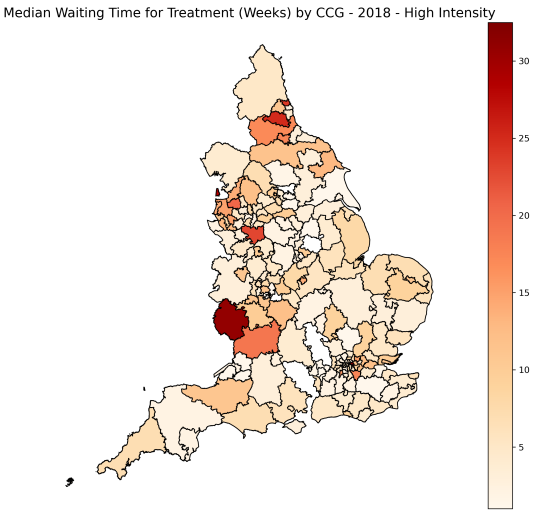
Figure C.VI: Median Waiting Times in Weeks for Treatment by Clinical Commissioning Groups (CCGs) and Treatment Intensity, 2018



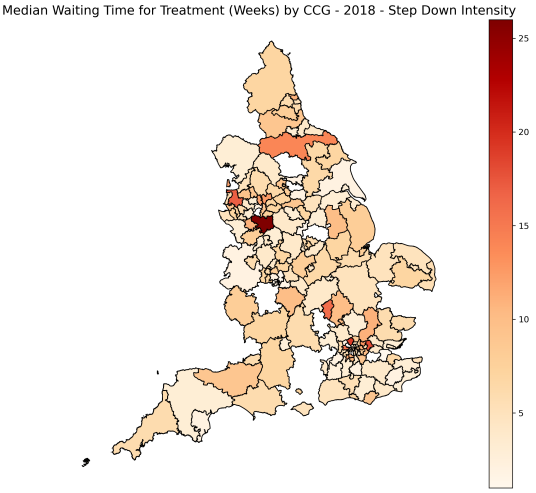
(a) All Intensities



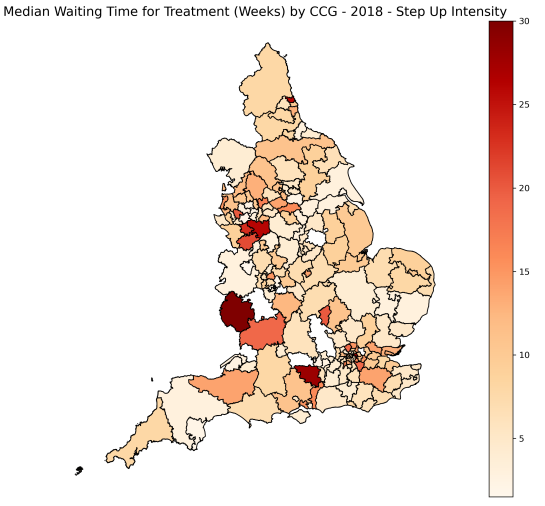
(b) Low Intensity



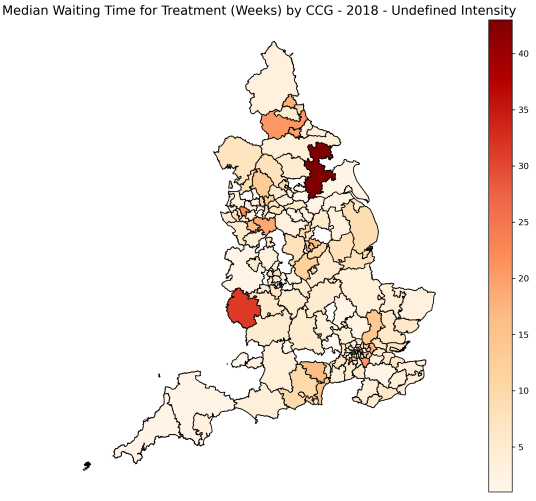
(c) High Intensity



(d) Step Down

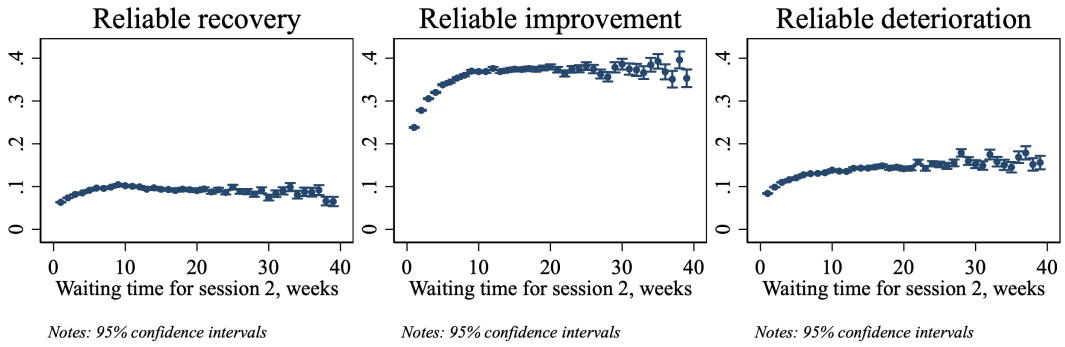


(e) Step Up



(f) Undefined

Figure C.VII: Main Outcomes for Different Waiting Times



Note: Own calculations.

D Balancing Properties

Table D.I: Balancing Properties of Covariates Between Default Treatment and 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
Therapy Controls								
Mental Health Index (Pre-Treatment)	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
Individual Controls								
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
Service Controls								
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
Local-Area Controls								
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 & Rubin, 2015; Imbens & Wooldridge, 2009).

Table D.II: Balancing Properties of Outcomes Between Default Treatment and 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 & Rubin, 2015; Imbens & Wooldridge, 2009).

E Summary Statistics for Treatment Durations and Outcomes by Waiting Times

Figure E.I: Histograms for Total Number of Sessions by Quartile of Waiting Time, All Treatments Intensities

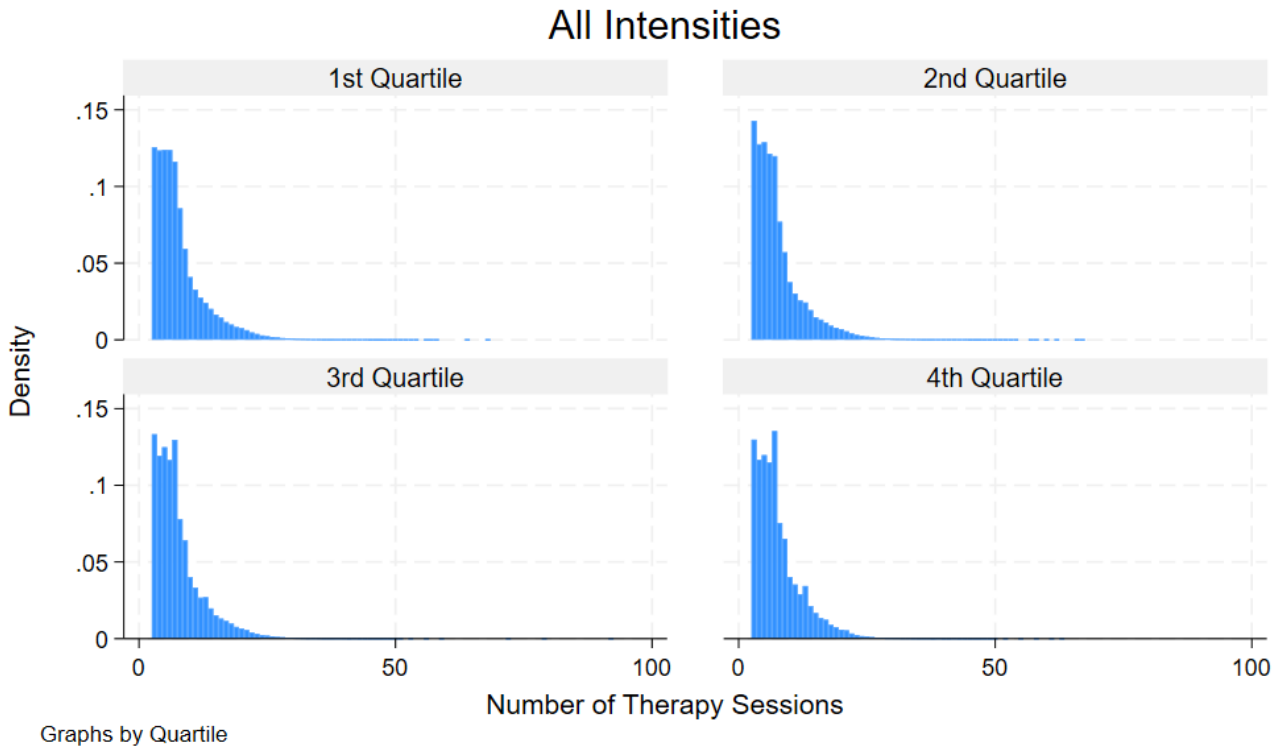


Table E.I: Summary Statistics for Total Number of Sessions by Quartile of Waiting Time, All Treatment Intensities

Quartile	Mean	SD
1	7.826	4.821
2	7.583	4.715
3	7.690	4.623
4	7.695	4.417

Note: The table shows the means and standard deviations of Total Number of Sessions by quartile of waiting time for all intensities combined. Quartiles represent the distribution of waiting times, with Quartile 1 being the shortest waiting times and Quartile 4 the longest.

Figure E.II: Histograms for Treatment Duration in Weeks by Quartile of Waiting Time, by Treatments Intensities

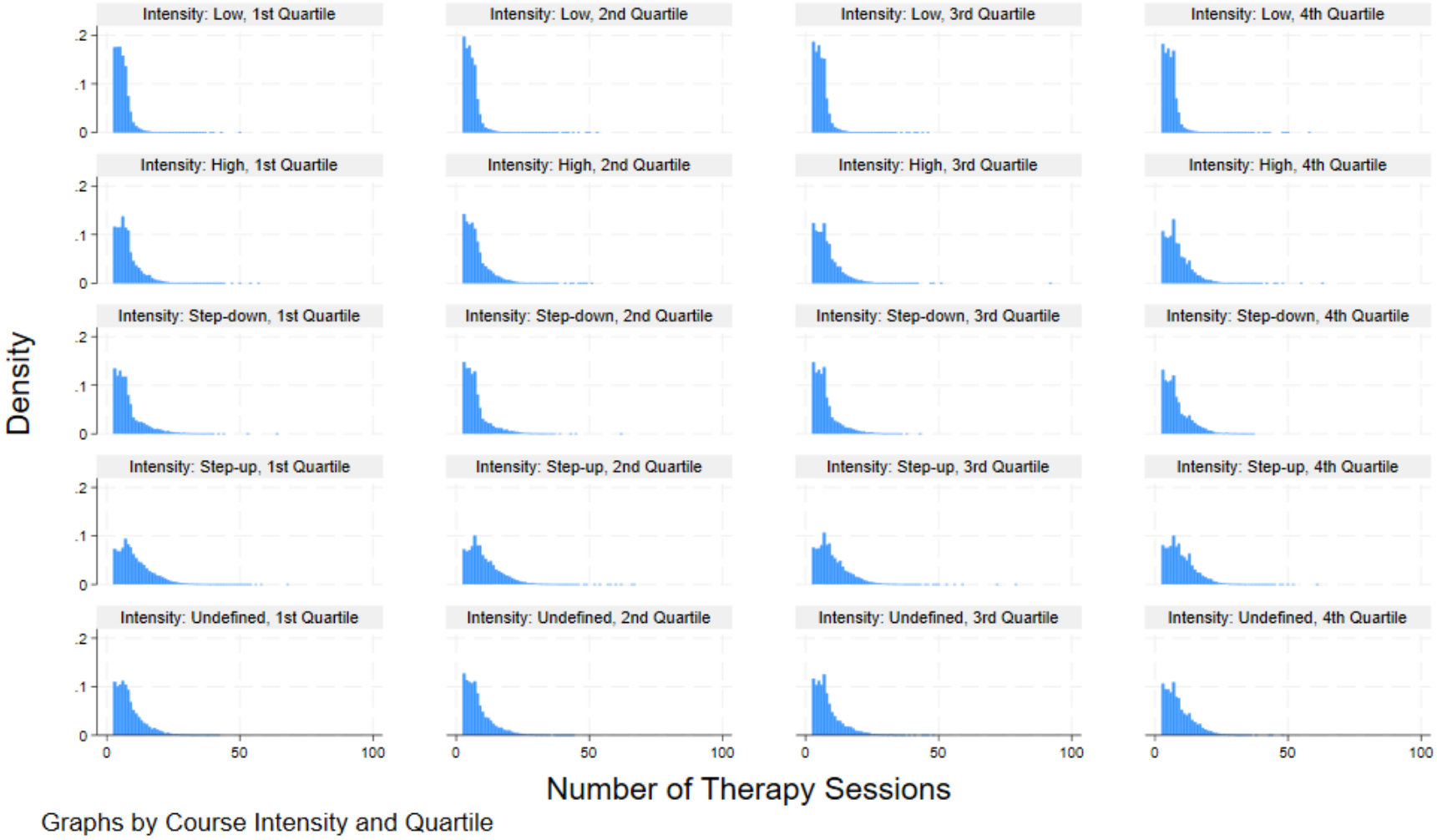


Table E.II: Summary Statistics for Total Number of Sessions by Quartile of Waiting Time and Treatment Intensity

Intensity	**Quartile 1**	**Quartile 2**	**Quartile 3**	**Quartile 4**
Low Intensity	5.751 (2.518)	5.627 (2.508)	5.698 (2.479)	5.716 (2.471)
High Intensity	7.616 (4.172)	7.468 (4.323)	7.967 (4.503)	8.351 (4.465)
Step Down	7.971 (5.240)	7.382 (4.706)	7.414 (4.570)	8.099 (4.744)
Step Up	10.405 (5.928)	10.229 (5.794)	9.854 (5.520)	9.622 (5.125)
Undefined	8.463 (5.002)	8.217 (5.078)	8.372 (5.134)	8.888 (5.083)

Note: The table shows the mean number of sessions (with standard deviations in parentheses) for each quartile of waiting time by treatment intensity. Quartiles represent waiting time distributions, with Quartile 1 being the shortest waiting times and Quartile 4 the longest.

Figure E.III: Histograms for Treatment Duration in Weeks by Quartile of Waiting Time, All Treatments Intensities

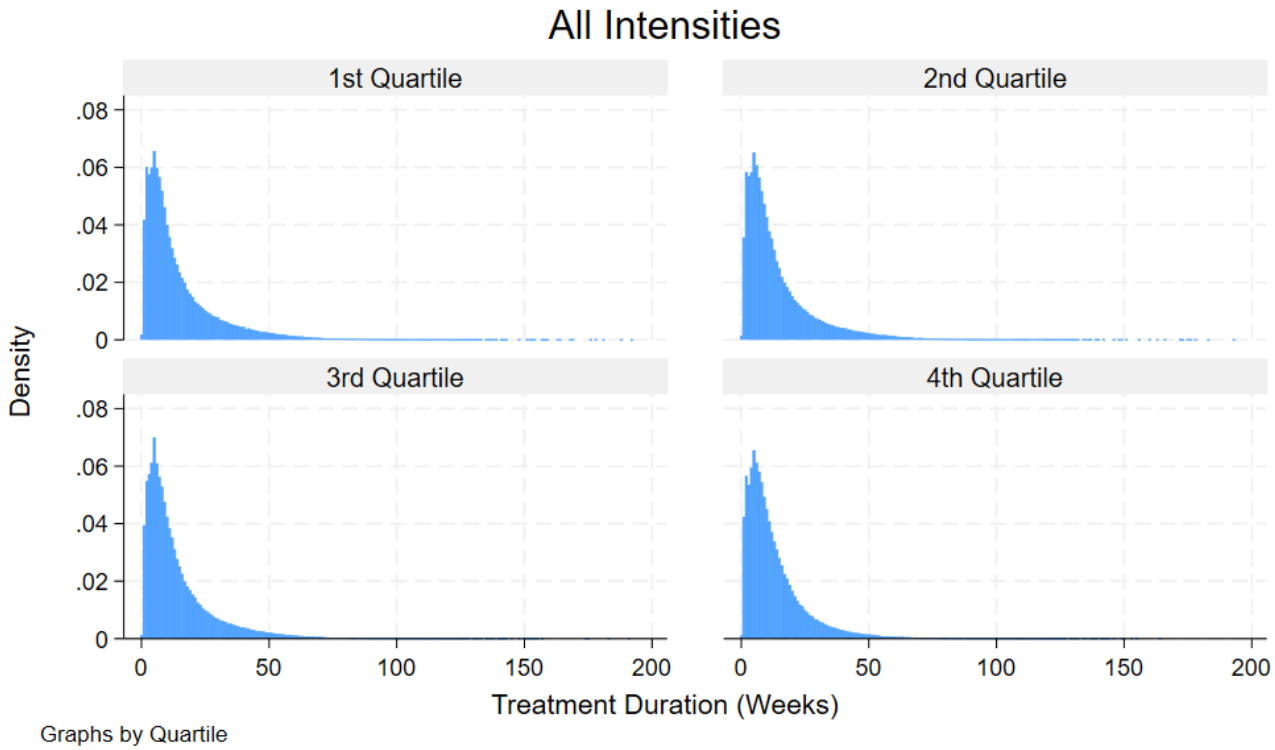


Table E.III: Summary Statistics for Treatment Duration in Weeks by Quartile of Waiting Time, All Treatment Intensities

Quartile	Mean	SD
1	14.520	14.563
2	14.415	14.257
3	13.986	13.777
4	12.986	12.016

Note: The table shows the means and standard deviations of Treatment Duration in Weeks by quartile of waiting time for all intensities combined. Quartiles represent the distribution of waiting times, with Quartile 1 being the shortest waiting times and Quartile 4 the longest.

Figure E.IV: Histograms for Treatment Duration in Weeks by Quartile of Waiting Time, by Treatments Intensities

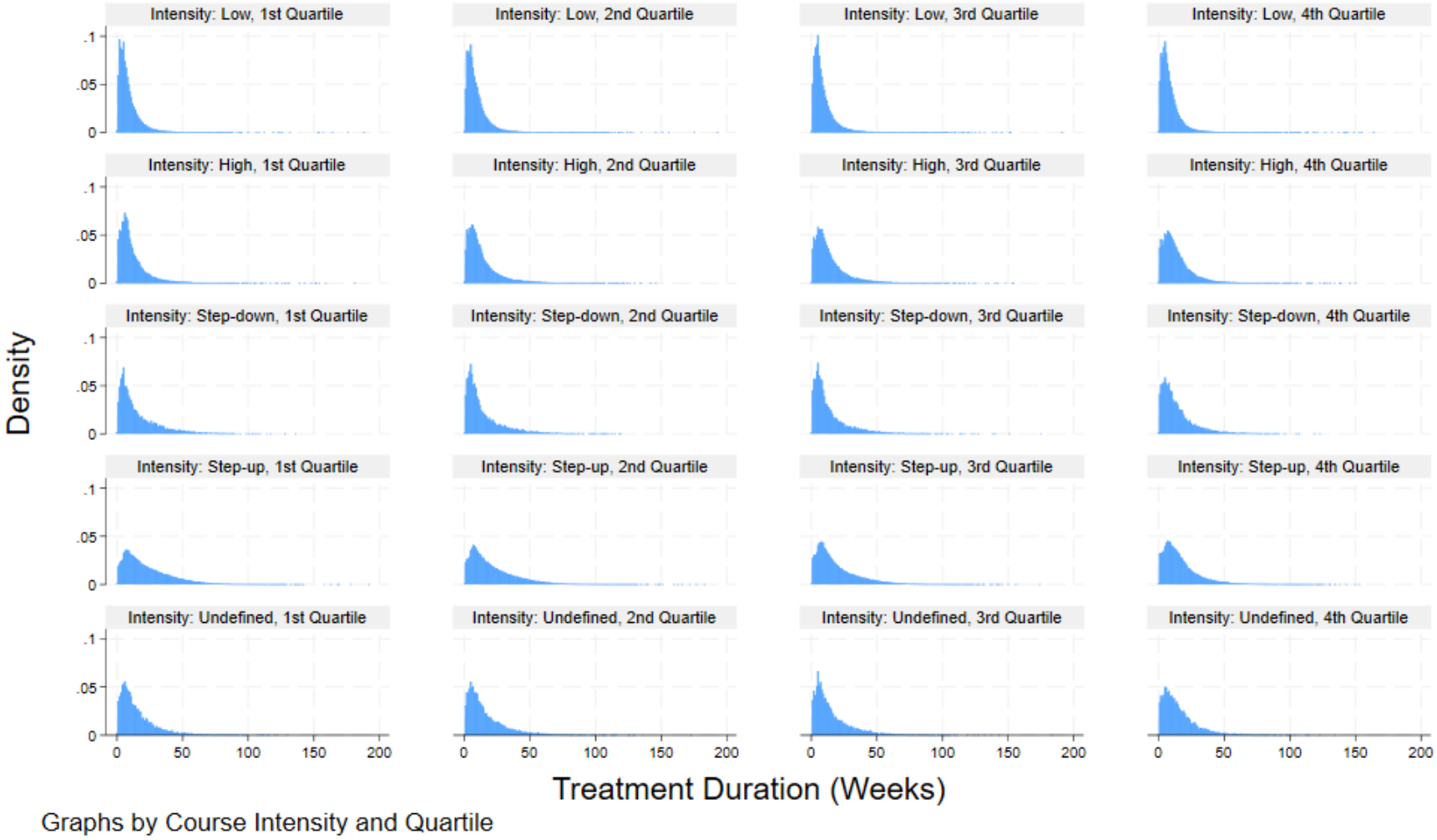


Table E.IV: Summary Statistics for Treatment Duration in Weeks by Quartile of Waiting Time and Treatment Intensity

Intensity	Quartile 1	Quartile 2	Quartile 3	Quartile 4
Low Intensity	8.890 (8.416)	9.465 (8.951)	9.267 (9.024)	8.954 (8.217)
High Intensity	12.084 (11.411)	13.743 (13.007)	14.292 (13.085)	13.998 (11.747)
Step Down	16.357 (15.889)	14.815 (14.970)	14.258 (14.687)	14.444 (13.283)
Step Up	22.684 (18.028)	21.219 (17.512)	19.218 (16.495)	16.997 (14.153)
Undefined	15.701 (14.614)	16.572 (15.376)	15.912 (15.105)	16.003 (13.837)

Note: The table shows the means (with standard deviations in parentheses) of treatment duration in weeks by quartile of waiting time and treatment intensity. Quartiles represent the distribution of waiting times, with Quartile 1 being the shortest waiting times and Quartile 4 the longest.

Figure E.V: Histograms for PHQ-9 Measured at the First Session by Quartile of Waiting Time, All Treatments Intensities.

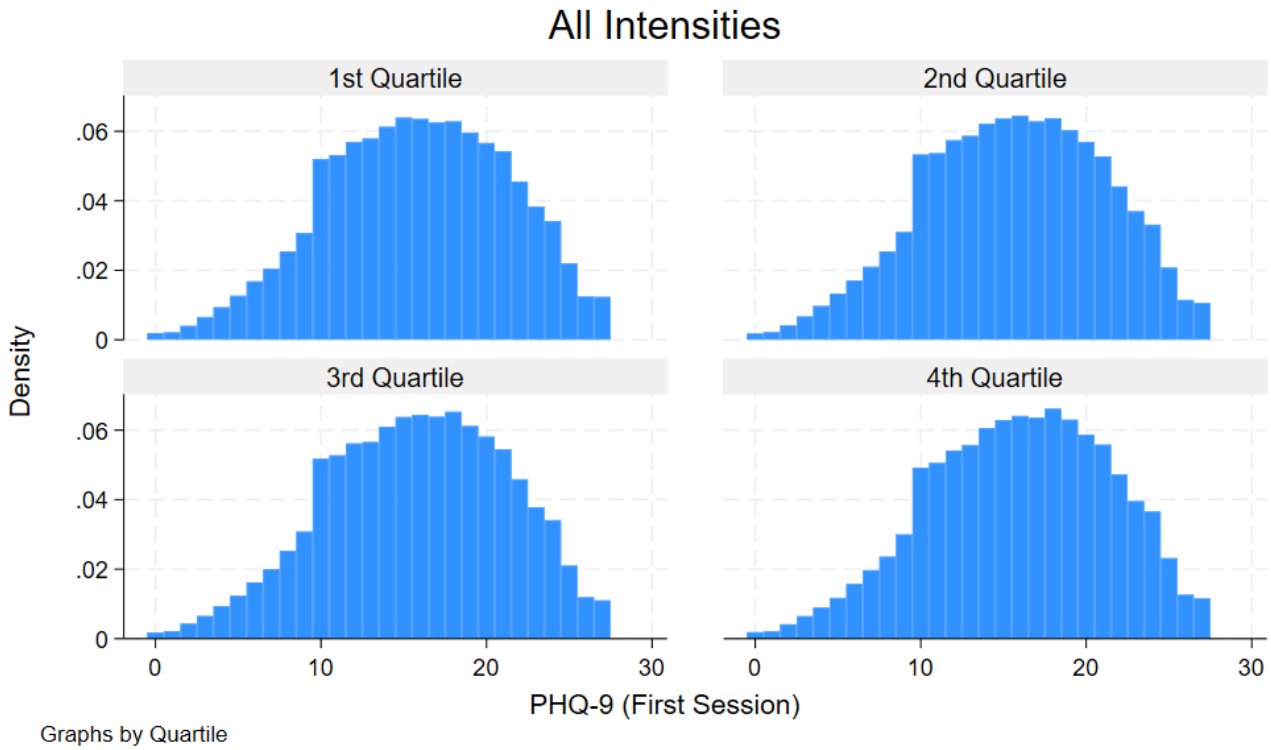


Table E.V: Summary Statistics for PHQ-9 Scores at First Session by Quartile of Waiting Time, All Intensities

Quartile	Mean	SD
1	15.744	5.520
2	15.636	5.488
3	15.755	5.482
4	15.923	5.497

Note: The table shows the means and standard deviations of PHQ-9 scores at the first session by quartile of waiting time for all intensities combined. Quartiles represent the distribution of waiting times, with Quartile 1 being the shortest waiting times and Quartile 4 the longest.

Figure E.VI: Histograms for PHQ-9 Measured at the First Session by Quartile of Waiting Time, by Treatments Intensities

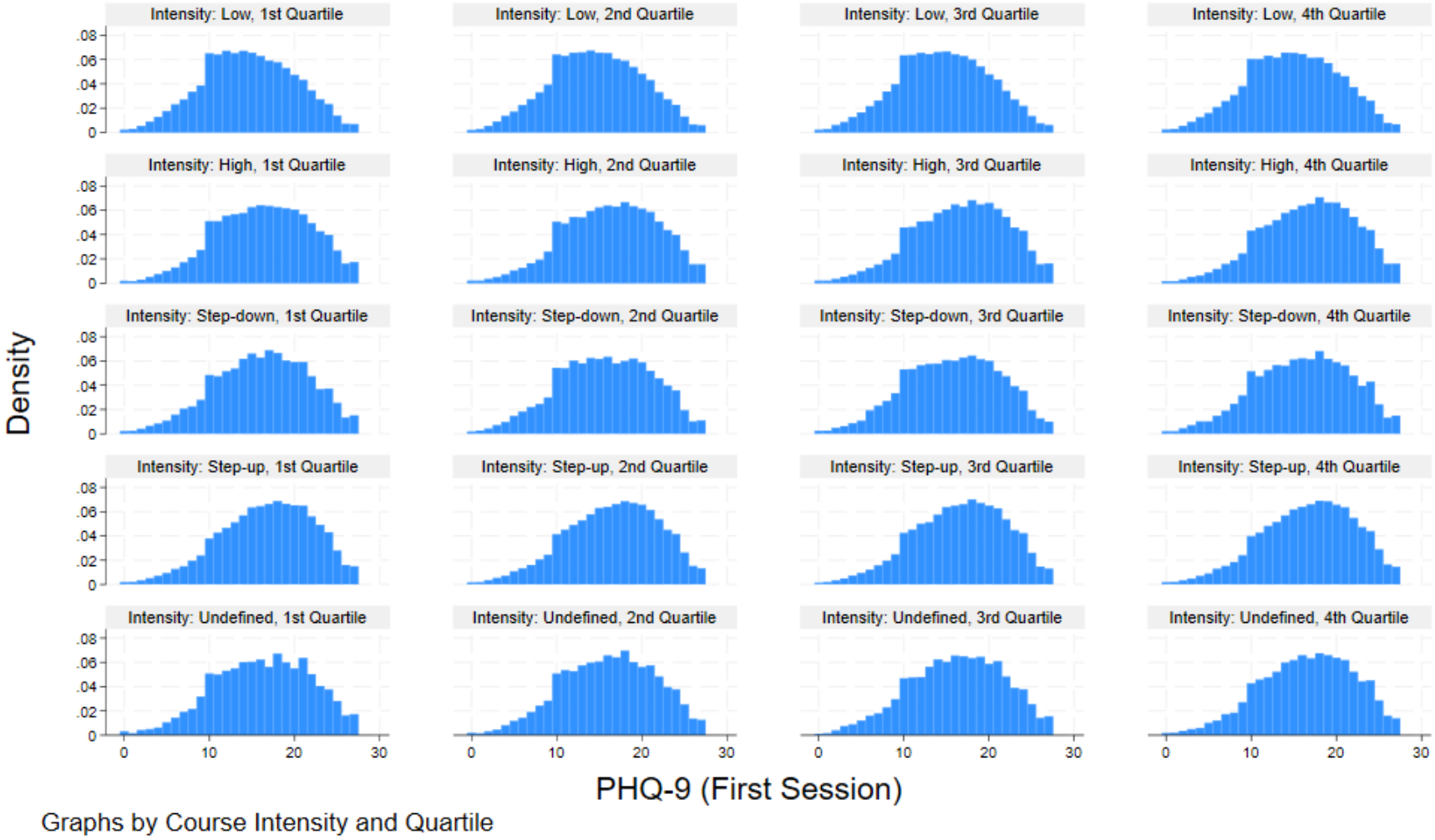


Table E.VI: Summary Statistics for PHQ-9 Scores at First Session by Quartile of Waiting Time and Treatment Intensity

Intensity	Quartile 1	Quartile 2	Quartile 3	Quartile 4
Low Intensity	14.596 (5.422)	14.585 (5.385)	14.628 (5.381)	14.882 (5.450)
High Intensity	16.275 (5.479)	16.334 (5.451)	16.544 (5.445)	16.659 (5.392)
Step Down	16.051 (5.518)	15.618 (5.538)	15.667 (5.547)	16.042 (5.581)
Step Up	16.692 (5.414)	16.478 (5.419)	16.515 (5.379)	16.677 (5.410)
Undefined	16.201 (5.568)	16.050 (5.442)	16.119 (5.529)	16.537 (5.435)

Note: The table shows the means (with standard deviations in parentheses) of PHQ-9 scores at the first session by quartile of waiting time and treatment intensity. Quartiles represent the distribution of waiting times, with Quartile 1 being the shortest waiting times and Quartile 4 the longest.

Figure E.VII: Histograms for PHQ-9 Measured at the Last Session by Quartile of Waiting Time, All Treatments Intensities

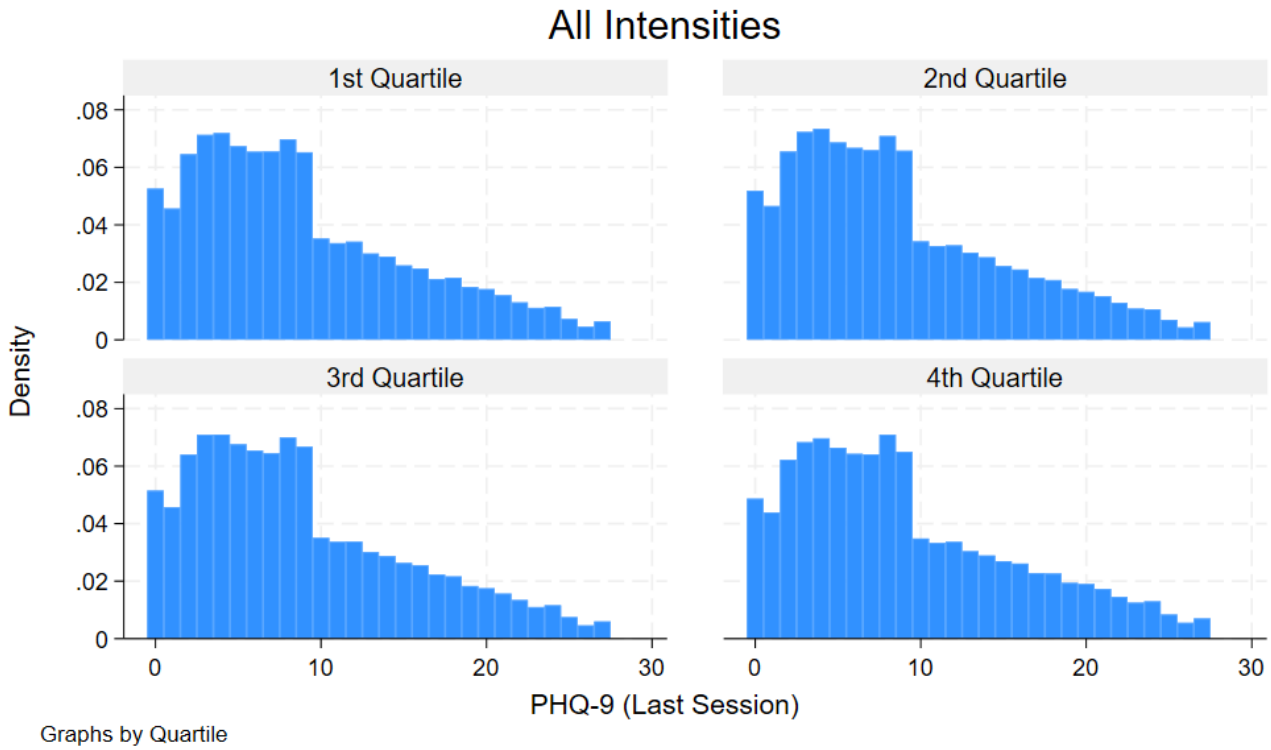
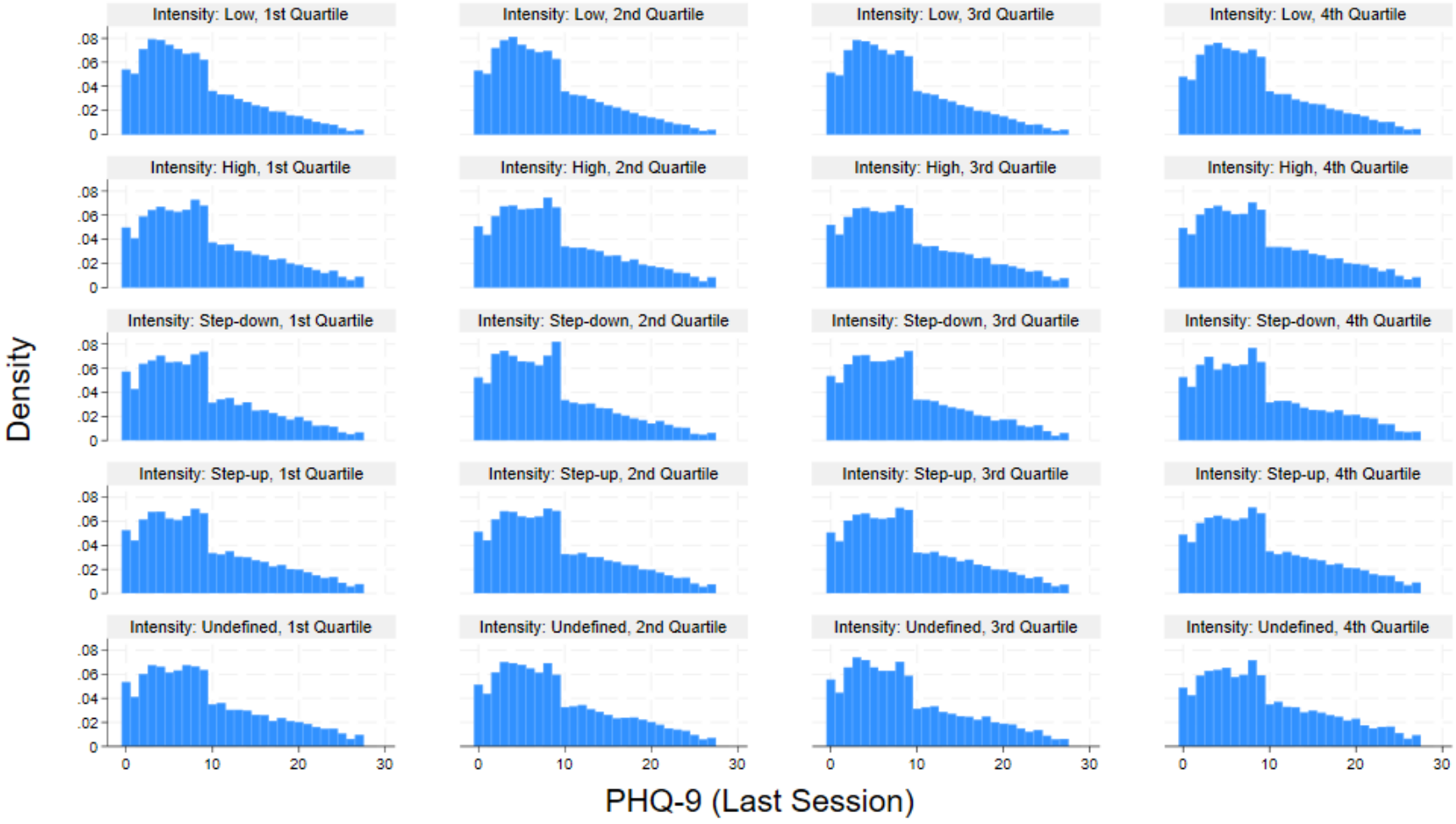


Table E.VII: Summary Statistics for PHQ-9 Scores at Last Session by Quartile of Waiting Time, All Intensities

Quartile	Mean	SD
1	8.788	6.482
2	8.690	6.428
3	8.826	6.481
4	9.083	6.617

Note: The table shows the means and standard deviations of PHQ-9 scores at the last session by quartile of waiting time for all intensities combined. Quartiles represent the distribution of waiting times, with Quartile 1 being the shortest waiting times and Quartile 4 the longest.

Figure E.VIII: Histograms for PHQ-9 Measured at the Last Session by Quartile of Waiting Time, by Treatments Intensities



XXX

Table E.VIII: Summary Statistics for PHQ-9 Scores at Last Session by Quartile of Waiting Time and Treatment Intensity

Intensity	Quartile 1	Quartile 2	Quartile 3	Quartile 4
Low Intensity	8.168 (6.122)	8.121 (6.082)	8.234 (6.112)	8.590 (6.307)
High Intensity	9.236 (6.632)	9.073 (6.598)	9.252 (6.697)	9.331 (6.778)
Step Down	8.861 (6.505)	8.555 (6.369)	8.742 (6.463)	9.230 (6.751)
Step Up	9.171 (6.698)	9.125 (6.656)	9.231 (6.676)	9.476 (6.805)
Undefined	9.326 (6.831)	9.145 (6.721)	8.947 (6.700)	9.572 (6.878)

Note: The table shows the means (with standard deviations in parentheses) of PHQ-9 scores at the last session by quartile of waiting time and treatment intensity. Quartiles represent the distribution of waiting times, with Quartile 1 being the shortest waiting times and Quartile 4 the longest.

Figure E.IX: Histograms for GAD-7 Measured at the First Session by Quartile of Waiting Time, All Treatments Intensities.

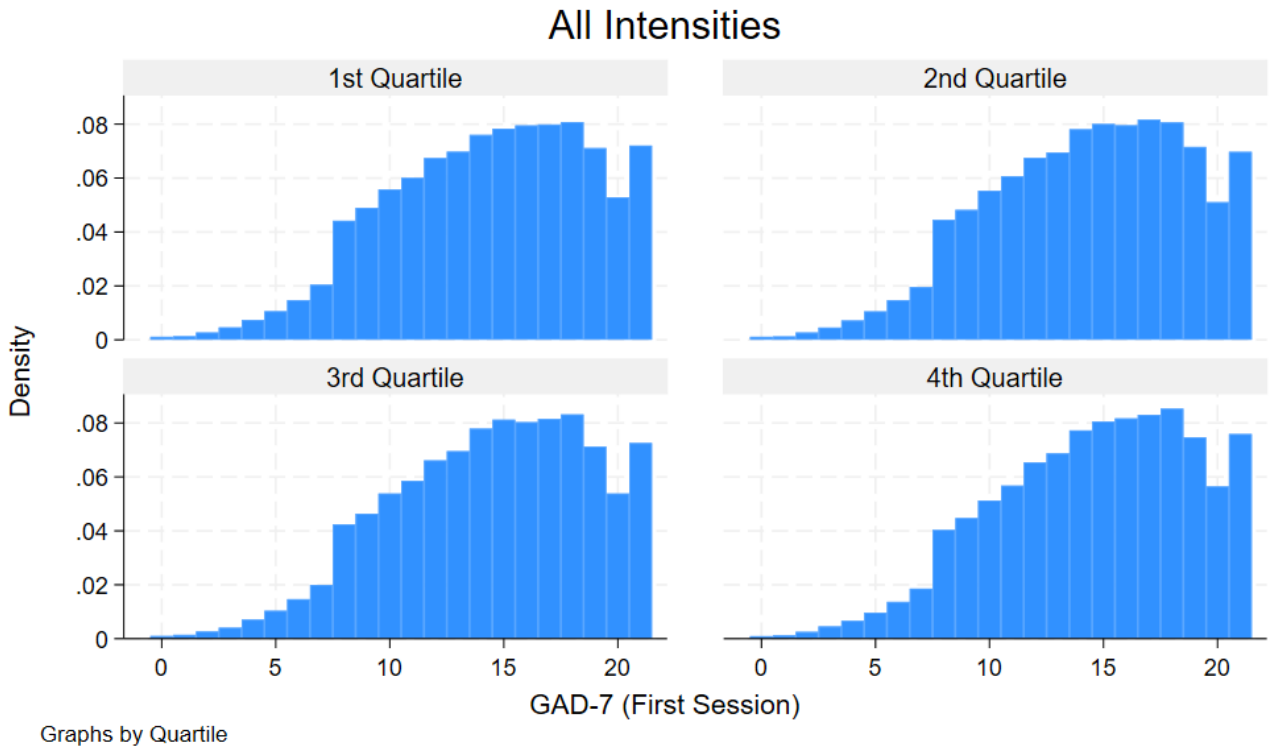


Table E.IX: Summary Statistics for GAD-7 Scores at First Session by Quartile of Waiting Time, All Intensities

Quartile	Mean	SD
1	14.312	4.366
2	14.307	4.336
3	14.397	4.332
4	14.536	4.315

Note: The table shows the means and standard deviations of GAD-7 scores at the first session by quartile of waiting time for all intensities combined. Quartiles represent the distribution of waiting times, with Quartile 1 being the shortest waiting times and Quartile 4 the longest.

Figure E.X: Histograms for GAD-7 Measured at the First Session by Quartile of Waiting Time, by Treatments Intensities

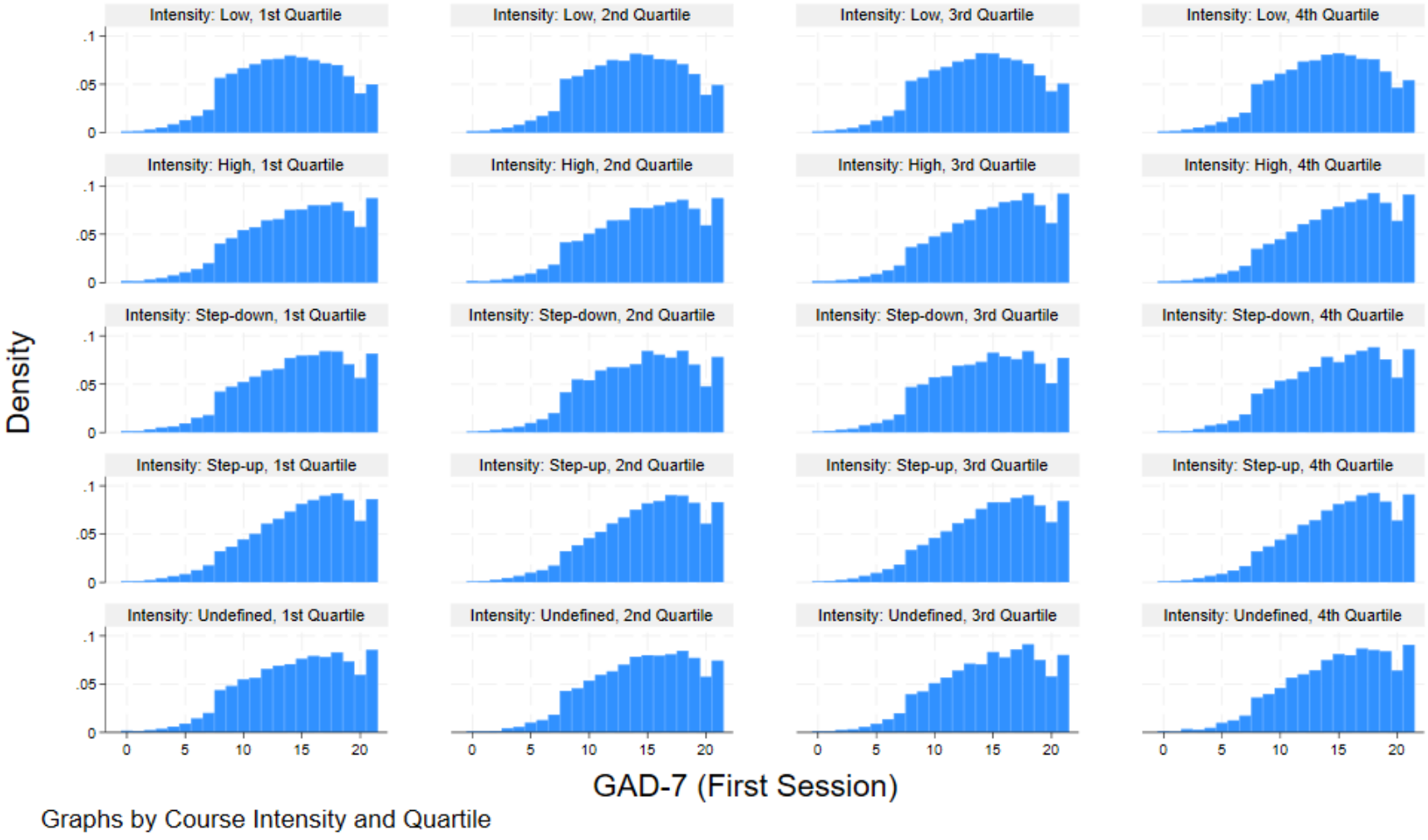


Table E.X: Summary Statistics for GAD-7 Scores at First Session by Quartile of Waiting Time and Treatment Intensity

Intensity	Quartile 1	Quartile 2	Quartile 3	Quartile 4
Low Intensity	13.634 (4.313)	13.698 (4.290)	13.754 (4.293)	13.948 (4.286)
High Intensity	14.539 (4.418)	14.660 (4.357)	14.868 (4.309)	14.926 (4.289)
Step Down	14.503 (4.375)	14.356 (4.340)	14.372 (4.341)	14.658 (4.316)
Step Up	14.936 (4.274)	14.830 (4.289)	14.808 (4.303)	14.975 (4.284)
Undefined	14.525 (4.394)	14.525 (4.289)	14.644 (4.293)	14.878 (4.300)

Note: The table shows the means (with standard deviations in parentheses) of GAD-7 scores at the first session by quartile of waiting time and treatment intensity. Quartiles represent the distribution of waiting times, with Quartile 1 being the shortest waiting times and Quartile 4 the longest.

Figure E.XI: Histograms for GAD-7 Measured at the Last Session by Quartile of Waiting Time, All Treatments Intensities.

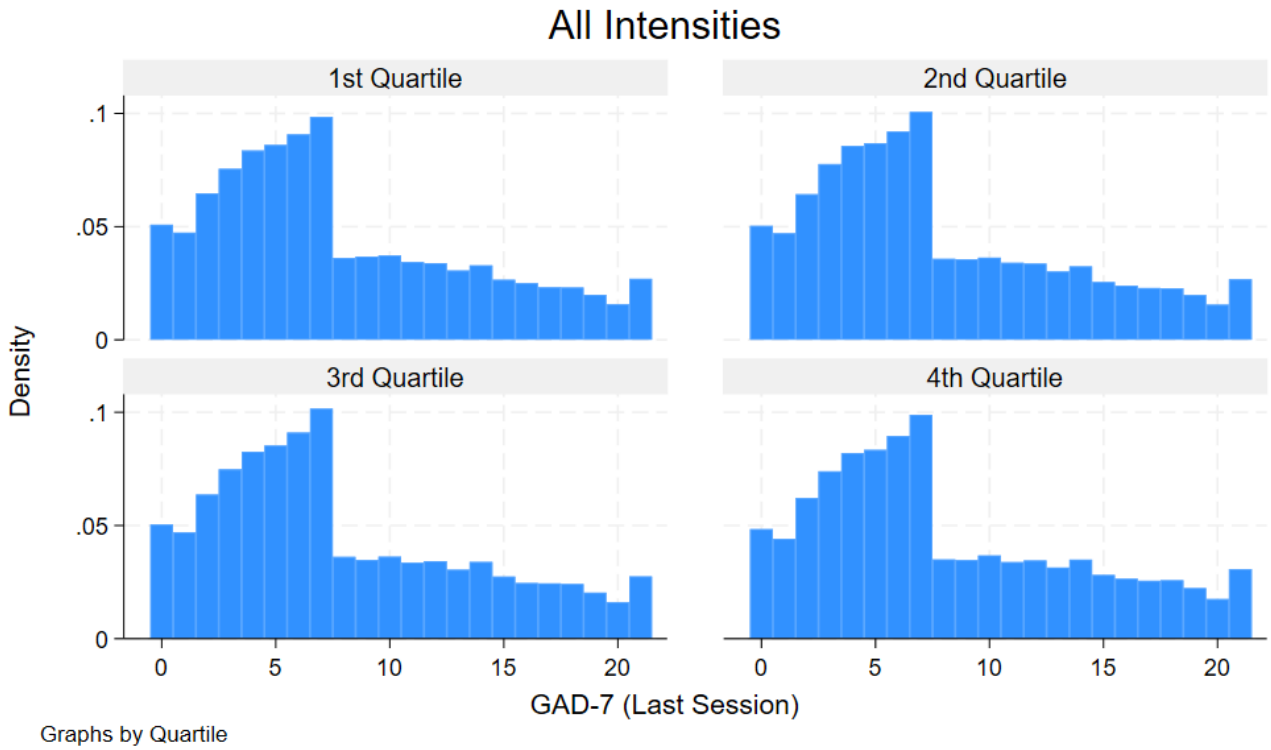


Table E.XI: Summary Statistics for GAD-7 Scores at Last Session by Quartile of Waiting Time, All Intensities

Quartile	Mean	SD
1	7.908	5.628
2	7.852	5.604
3	7.967	5.655
4	8.172	5.749

Note: The table shows the means and standard deviations of GAD-7 scores at the last session by quartile of waiting time for all intensities combined. Quartiles represent the distribution of waiting times, with Quartile 1 being the shortest waiting times and Quartile 4 the longest.

Figure E.XII: Histograms for GAD-7 Measured at the Last Session by Quartile of Waiting Time, by Treatments Intensities

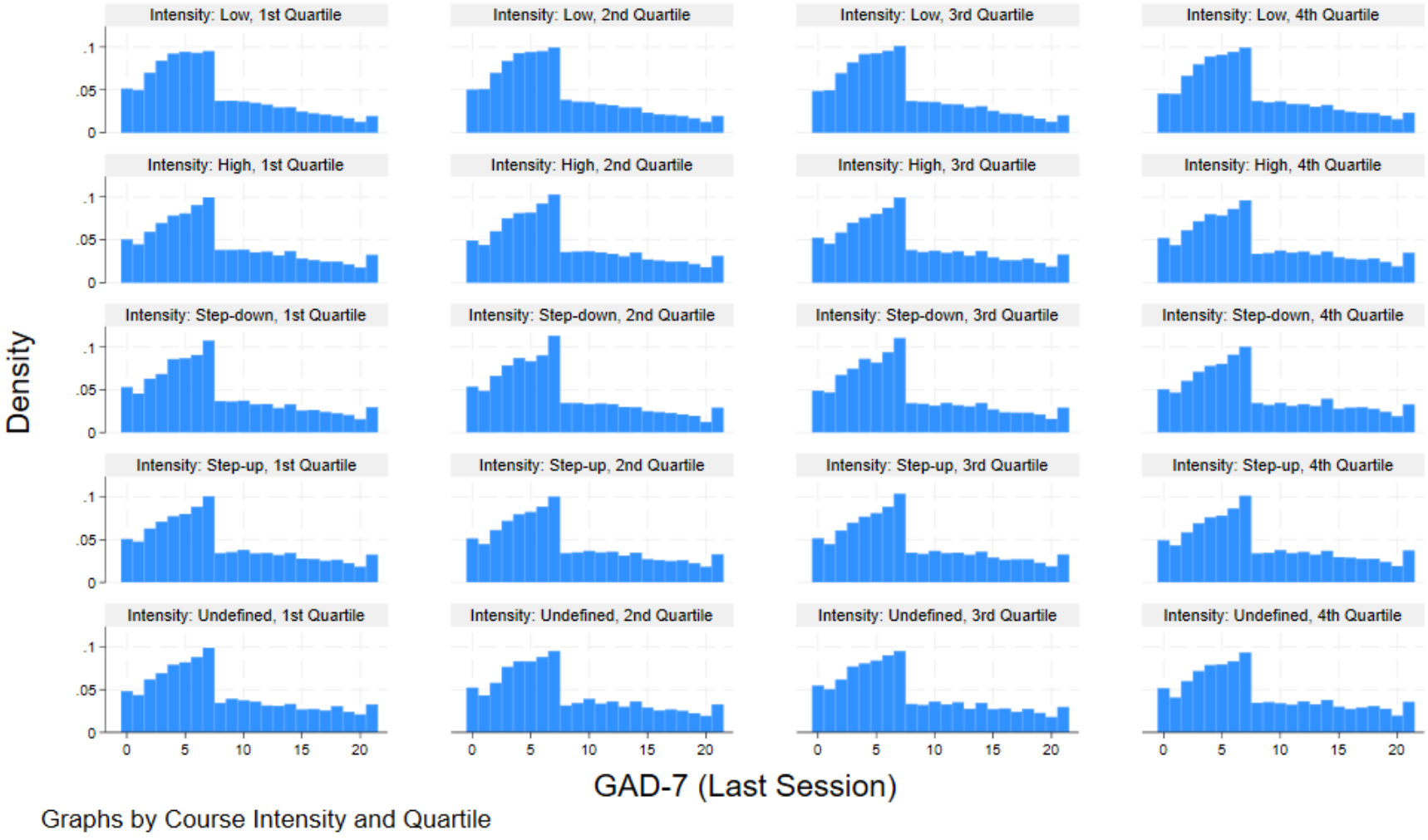


Table E.XII: Summary Statistics for GAD-7 Scores at Last Session by Quartile of Waiting Time and Treatment Intensity

Intensity	Quartile 1	Quartile 2	Quartile 3	Quartile 4
Low Intensity	7.440 (5.367)	7.407 (5.346)	7.508 (5.379)	7.801 (5.522)
High Intensity	8.222 (5.730)	8.134 (5.718)	8.303 (5.810)	8.352 (5.869)
Step Down	7.941 (5.639)	7.730 (5.581)	7.917 (5.652)	8.318 (5.864)
Step Up	8.211 (5.801)	8.206 (5.787)	8.276 (5.809)	8.473 (5.888)
Undefined	8.335 (5.839)	8.225 (5.815)	8.046 (5.803)	8.508 (5.946)

Note: The table shows the means (with standard deviations in parentheses) of GAD-7 scores at the last session by quartile of waiting time and treatment intensity. Quartiles represent the distribution of waiting times, with Quartile 1 being the shortest waiting times and Quartile 4 the longest.

Table E.XIII: A Model for Waiting Time Between Initial Assessment and First Clinical Session in Weeks

	All Intensities	Low Int.	High Int.	Step Up	Step Down	Not Recorded
	(1)	(2)	(3)	(4)	(5)	(6)
Mental Health Index,	0.118***	0.130***	0.123***	0.035	-0.031	0.275***
Pre-Treatment (Z-Score)	(0.029)	(0.032)	(0.033)	(0.046)	(0.071)	(0.076)
Number of Individuals	1,246,792	491,942	275,990	388,136	44,396	46,328
R Squared	0.219	0.197	0.285	0.231	0.201	0.239
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

Note: Robust standard errors clustered at the clinical-commissioning-group level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

F Additional Results

Table F.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 F.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 F.III: 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 F.IV: 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.

G Robustness Checks

Table G.I: Robustness Check: Selection on Outcome – Redefining Treatment Completion

	Reliable Recovery (0-1)		Reliable Improvement (0-1)		Reliable Deterioration (0-1)	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: End of Treatment = Last Clinical Session - 1						
Treatment	0.267***	0.269***	0.325***	0.324***	-0.076***	-0.075***
	(0.004)	(0.004)	(0.003)	(0.003)	(0.002)	(0.001)
Number of Individuals	1,163,182	1,163,182	1,163,182	1,163,182	1,163,182	1,163,182
Treatment Group	535,881	535,881	535,881	535,881	535,881	535,881
Control Group	627,301	627,301	627,301	627,301	627,301	627,301
R Squared	0.105	0.170	0.105	0.145	0.016	0.065
Panel B: End of Treatment = Last Clinical Session - 2						
Treatment	0.196***	0.199***	0.281***	0.283***	-0.068***	-0.067***
	(0.004)	(0.003)	(0.004)	(0.003)	(0.002)	(0.001)
Number of Individuals	1,085,231	1,085,231	1,085,231	1,085,231	1,085,231	1,085,231
Treatment Group	457,930	457,930	457,930	457,930	457,930	457,930
Control Group	627,301	627,301	627,301	627,301	627,301	627,301
R Squared	0.065	0.129	0.077	0.118	0.012	0.065
Individual Controls	No	Yes	No	Yes	No	Yes
Therapy Controls	No	Yes	No	Yes	No	Yes
Local-Area Controls	No	Yes	No	Yes	No	Yes
Local-Area 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 G.II: Robustness Check: Selection on Outcome – Grouping Treatment Intensities

	Reliable Recovery (0-1)		Reliable Improvement (0-1)		Reliable Deterioration (0-1)	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Low Intensity + Step Up						
Treatment	0.445***	0.431***	0.385***	0.370***	-0.086***	-0.083***
	(0.004)	(0.004)	(0.003)	(0.003)	(0.002)	(0.001)
Number of Individuals	880,078	880,078	880,078	880,078	880,078	880,078
Treatment Group	432,092	432,092	432,092	432,092	432,092	432,092
Control Group	447,986	447,986	447,986	447,986	447,986	447,986
R Squared	0.229	0.290	0.149	0.186	0.022	0.064
Panel B: High Intensity + Step Down						
Treatment	0.440***	0.432***	0.400***	0.394***	-0.085***	-0.085***
	(0.007)	(0.007)	(0.006)	(0.005)	(0.003)	(0.002)
Number of Individuals	320,386	320,386	320,386	320,386	320,386	320,386
Treatment Group	163,955	163,955	163,955	163,955	163,955	163,955
Control Group	156,431	156,431	156,431	156,431	156,431	156,431
R Squared	0.228	0.292	0.162	0.196	0.022	0.068
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 G.III: Average Treatment Effects on Mental Health - Additionally Controlling for Total Number of Sessions and Total Treatment Duration

	Reliable Recovery (0-1)	Reliable Improvement (0-1)	Reliable Deterioration (0-1)
	(1)	(2)	(3)
Treatment	0.430***	0.372***	-0.083***
	(0.004)	(0.003)	(0.001)
Number of Individuals	1,246,792	1,246,792	1,246,792
Treatment Group	618,574	618,574	618,574
Control Group	628,218	628,218	628,218
R Squared	0.292	0.191	0.065
Individual Controls	Yes	Yes	Yes
Therapy Controls	Yes	Yes	Yes
Local-Area Controls	Yes	Yes	Yes
Local-Area Fixed Effects	Yes	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes

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

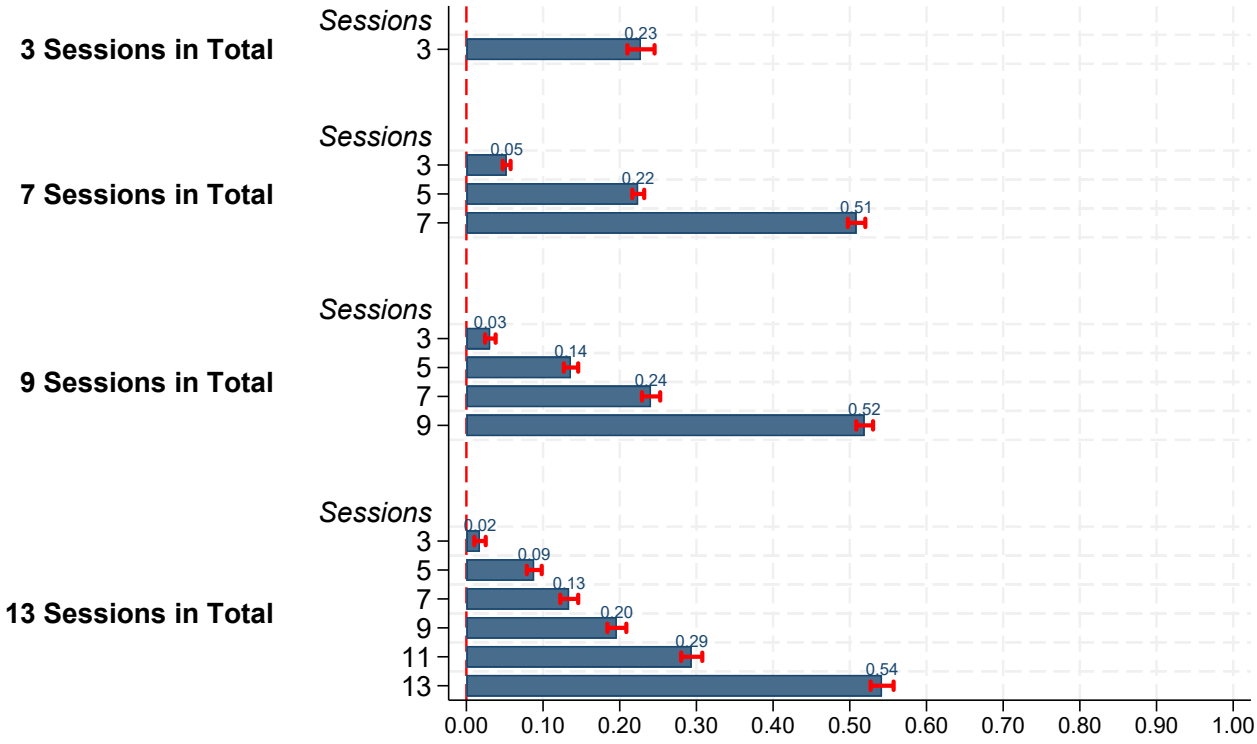


Figure G.I: Reliable Recovery – Session Value Added

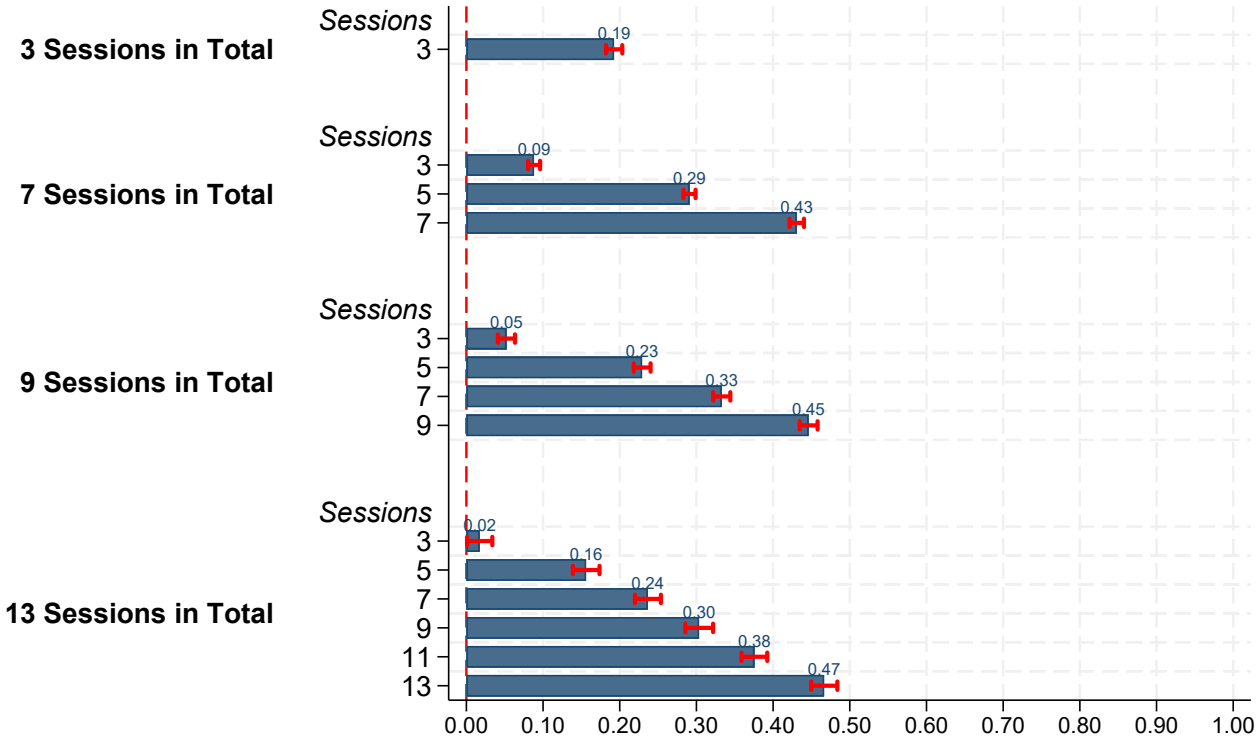


Figure G.II: Reliable Improvement – Session Value Added

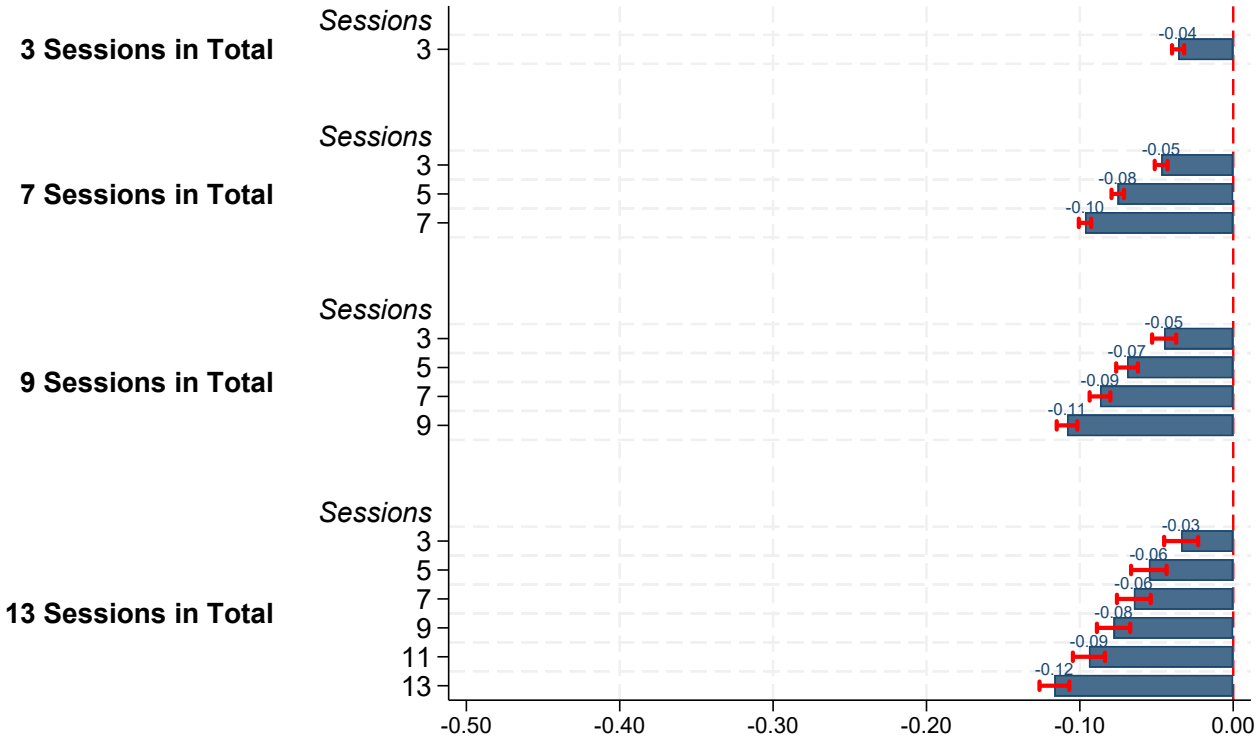


Figure G.III: Reliable Deterioration – Session Value Added

Table G.IV: Average Treatment Effects on Mental Health - Additionally Controlling for Session Spacing

	Reliable Recovery (0-1)		Reliable Improvement (0-1)		Reliable Deterioration (0-1)	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.431***	0.430***	0.377***	0.376***	-0.084***	-0.084***
	(0.004)	(0.004)	(0.003)	(0.003)	(0.001)	(0.001)
Weeks Per Session	No	Yes	No	Yes	No	Yes
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	1,246,792	1,246,792	1,246,792	1,246,792	1,246,792	1,246,792
Treatment Group	619,491	619,491	619,491	619,491	619,491	619,491
Control Group	627,301	627,301	627,301	627,301	627,301	627,301
R Squared	0.289	0.290	0.187	0.189	0.064	0.064

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

Table G.V: Average Treatment Effects: Number of Weeks Between Sessions (Session Spacing)

	Reliable Recovery (1)	Reliable Improvement (2)	Reliable Deterioration (3)
≤25th Percentile (1.1 Weeks)			
Treatment	0.402***	0.356***	-0.077***
	(0.005)	(0.004)	(0.001)
Number of Individuals	311,675	311,675	311,675
Treatment Group	153,032	153,032	153,032
Control Group	158,643	158,643	158,643
R Squared	0.249	0.140	0.056
≥75th Percentile (2.4 Weeks)			
Treatment	0.391***	0.343***	-0.082***
	(0.006)	(0.005)	(0.002)
Number of Individuals	311,884	311,884	311,884
Treatment Group	166,638	166,638	166,638
Control Group	145,246	145,246	145,246
R Squared	0.284	0.200	0.071
≥90th Percentile (3.5 Weeks)			
Treatment	0.330***	0.291***	-0.070***
	(0.009)	(0.008)	(0.003)
Number of Individuals	128,479	128,479	128,479
Treatment Group	70,949	70,949	70,949
Control Group	57,530	57,530	57,530
R Squared	0.260	0.178	0.075
Individual Controls	Yes	Yes	Yes
Therapy Controls	Yes	Yes	Yes
Local-Area Controls	Yes	Yes	Yes
Local-Area Fixed Effects	Yes	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes

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

Table G.VI: Predicting Repeat Enrolment From Weeks on Waitlist Amongst Control-Group Patients

	Repeat Enrolment (0-1)	
	(1)	(2)
Weeks on Waitlist	-0.001***	-0.001***
	(0.000)	(0.000)
Individual Controls	No	Yes
Therapy Controls	No	Yes
Local-Area Controls	No	Yes
Local-Area Fixed Effects	No	Yes
Time Fixed Effects	No	Yes
Number of Individuals	627,301	627,301
Treatment Group	0	0
Control Group	627,301	627,301
R Squared	0.000	0.069

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

Table G.VII: Average Treatment Effects: Robustness – Excluding Repeat Enrolments

	Reliable Recovery (0-1)		Reliable Improvement (0-1)		Reliable Deterioration (0-1)	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.454***	0.441***	0.391***	0.380***	-0.087***	-0.085***
	(0.004)	(0.004)	(0.004)	(0.003)	(0.002)	(0.001)
Individual Controls	No	Yes	No	Yes	No	Yes
Therapy Controls	No	Yes	No	Yes	No	Yes
Local-Area Controls	No	Yes	No	Yes	No	Yes
Local-Area Fixed Effects	No	Yes	No	Yes	No	Yes
Time Fixed Effects	No	Yes	No	Yes	No	Yes
Number of Individuals	1,059,644	1,059,644	1,059,644	1,059,644	1,059,644	1,059,644
Treatment Group	518,366	518,366	518,366	518,366	518,366	518,366
Control Group	541,278	541,278	541,278	541,278	541,278	541,278
R Squared	0.237	0.299	0.155	0.19	0.022	0.064

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

Table G.VIII: 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 G.IX: 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.

H 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.[55]

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 H.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

[55] 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 H.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. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

I Heterogeneous Treatment Effects

Table I.I: Summary statistics 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 I.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 I.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 I.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 I.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